



**UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACION**

**OBTENCION DE UN CLASIFICADOR OPTIMO PARA LA EVALUACION DE
LA CALIDAD DE MODELOS TRIDIMENSIONALES DE PROTEINAS**

**TESIS PARA OPTAR AL GRADO DE MAGISTER EN CIENCIAS MENCION
COMPUTACION**

ISMAEL ALFONSO VERGARA CORREA

**PROFESOR GUIA:
GONZALO NAVARRO BADINO
FRANCISCO MELO LEDERMANN**

**MIEMBROS DE LA COMISION:
CARLOS HURTADO LARRAIN
JULIO CABALLERO RUIZ**

**SANTIAGO DE CHILE
JUNIO 2008**

Dedicatoria

*A mi esposa Gabriela,
quien hizo del amor
el axioma fundamental de mi vida.*

Agradecimientos

Qué bonita instancia esta, la de agradecer. La sola palabra me llena de recuerdos y emociones, e inevitablemente muchas caras se vienen a mi mente. No puedo evitar tampoco el reflexionar sobre el sentido que tiene el agradecer antes de escribir las siguientes líneas; la conclusión, bastante simple por lo demás, es que cada una de las personas a las que quiero agradecer tiene un don que compartió conmigo en una etapa crucial de mi vida, y que me ayudó a lograr un objetivo. Cada uno de ellos compartió su "gracia" conmigo.

Entonces, en un humilde intento por retribuirlos ante tamaña generosidad, dejo a su disposición mis gracias, para cuando las necesiten.

A Rodolfo Vergara del Pozo, por su fortaleza.

A Marcela Correa, por su amor incondicional de madre.

A mis hermanos, Isabel y Rodolfo, por su ejemplo.

A Gabriela, por su belleza.

A Francisco Melo, por su inteligencia.

A Gonzalo Navarro, por su paciencia y buena disposición.

A Angélica Aguirre, por su paciencia y buena disposición.

A Evandro Ferrada, por su amistad.

A mis compañeros del Melolab, por su alegría.

A mis compañeros BT y mis compañeros computines, por su compañía.

Y a FONDECYT #1051112, por su generosidad.

Indice de Contenido

RESUMEN	8
1) Introducción	9
1.1) Biología, el dogma central de la biología molecular, y la bioinformática	9
1.2) Relación secuencia-estructura en proteínas. Relación estructura-función en proteínas.	12
1.3) La necesidad de la predicción computacional de la estructura tridimensional de proteínas.	14
1.4) Métodos de predicción computacional de la estructura tridimensional de proteínas.	14
1.5) Errores posibles en el modelado comparativo de proteínas.	17
1.6) Minería de Datos	18
1.7) Objetivos	19
2) Marco Teórico	21
2.1) Generación de la base de datos.	21
2.2) Descripción de las variables medidas sobre los modelos.....	24
2.3) Estadística descriptiva de los datos y detección de outliers.....	32
2.4) Métodos de ranking de variables.....	35
2.5) Métodos de selección de variables.	38
2.6) Métodos de extracción de variables.	44
2.7) Algoritmos de aprendizaje y clasificación.	47
2.8) Uso de la base de datos para la obtención de clasificadores.	71
2.9) Medidas de rendimiento de clasificadores.	73
2.10) Test estadísticos de comparación de clasificadores	74
2.11) Software	77
3) Resultados	78
3.2) Medidas básicas y detección de outliers.	82
3.3) Rankers.....	84
3.4) Selección de variables	85
3.5) Análisis de componentes principales (ACP).....	87
3.6) Clasificadores.....	89
4) Comparación de clasificadores	100
4.1) Test de McNemar.	100
4.2) Test de Delong	103
4.3) Análisis sobre modelos incompletos.....	106
5) Discusión	108
5.1) Análisis estadístico básico de las variables utilizadas.....	108
5.2) Ranking de las variables.....	109
5.3) Selección de variables	110
5.4) Análisis de Componentes Principales	111
5.5) Variables como clasificadores.....	112
5.6) Clasificadores multivariados	113
5.7) Comparación de clasificadores.....	120
5.8) Consideración del espacio de modelos incompletos	121
6) Conclusiones.....	122
7) Referencias	123
8) Anexos.....	129
Anexo A: gráficos de probabilidad normal	129
Anexo B: gráficos de caja y bigotes	139

Anexo C: comportamiento de los índices de ranking medidos sobre las variables..	149
Anexo D: cantidad de información capturada por cada componente principal y tabla de pesos de las componentes principales.....	151
Anexo E: gráficos en el espacio de parámetros de SVM	155
Anexo F: valores de $\mathbf{y}_i \boldsymbol{\alpha}_i^*$ con su vector de soporte asociado \mathbf{x}_i	158
Anexo G: Clasificadores optimizados para MLP	174
Anexo H: intervalos de confianza de la diferencia de las AUC para cada par de variables como clasificadores.....	179
Anexo I: intervalos de confianza de la diferencia de las AUC para cada par de clasificadores multivariantes.	186

Índice de Figuras

Fig 1. Los pinzones de Darwin.	9
Fig 2. Molécula de ADN.	10
Fig 3. Dogma Central de la Biología Molecular.	11
Fig 4. Plegamiento de una proteína.	12
Fig 5. Relación secuencia – estructura en proteínas.	13
Fig 6. Modelado por homología.	16
Fig 7. Errores en el modelado por homología.	18
Fig 8. Estructura general de un aminoácido.	22
Fig 9. Esquema de la base de datos.	24
Fig 10. Categorías utilizadas en la medición de variables.	25
Fig 11. Propiedades del alineamiento secuencia–estructura.	29
Fig 12. Propiedades estructurales medidas sobre los modelos.	32
Fig 13. Orden en el espacio de búsqueda de variables.	39
Fig 14. Componentes principales sobre los datos.	45
Fig 15. Esquema de la distribución de modelos correctos e incorrectos para una variable.	48
Fig 16. Esquema de la distribución de modelos correctos e incorrectos en el espacio n-dimensional.	49
Fig 17. Esquema de discretización del rango de valores de una variable continua.	51
Fig 18. Múltiples hiperplanos separadores.	52
Fig 19. Hiperplano equidistante y margen geométrico.	53
Fig 20. Variables de holgura en el problema de optimización.	57
Fig 21. Ejemplo de árbol de decisión.	59
Fig 22. Diagrama de flujo de un algoritmo genético general.	60
Fig 23. Arquitectura de un perceptrón.	63
Fig 24. Arquitectura de un perceptrón multicapa.	64
Fig 25. Control del sobreajuste en un perceptrón multicapa.	67
Fig 26. Validación cruzada de orden k.	72
Fig 27. Curvas ROC y AUC.	74
Fig 28. El problema enfrentado en este trabajo.	78
Fig 29. Esquema de la primera etapa de este trabajo.	80
Fig 30. Esquema de la segunda etapa de este trabajo.	81
Fig 31. Porcentaje de calce entre los distintos métodos de ranking.	85
Fig 32. Variables en la primera y segunda componente principal.	88
Fig 33. Variables en la tercera y cuarta componente principal.	88
Fig 34. Varianza capturada por componente principal, y su distribución acumulada.	89
Fig 35. Frecuencia de ocurrencia de cada variable en las fórmulas de GA Math. .	93
Fig 36. Frecuencia de ocurrencia de cada variable en las reglas lógicas de GA Logic.	95
Fig 37. Arquitectura del MLP óptimo con una capa oculta.	97
Fig 38. Curvas ROC para los clasificadores multivariables y los mejores cinco clasificadores univariables.	103
Fig 39. Distribución acumulada de la variable Target Coverage.	106

Índice de Tablas

Tabla 1. Definición de estadísticos básicos.	33
Tabla 2. Espacio de valores posibles para los parámetros de SVM.....	58
Tabla 3. Espacio de valores muestreados para los parámetros de SVM.	58
Tabla 4. Parámetros utilizados en la ejecución de GA Logic.	61
Tabla 5. Parámetros utilizados en la ejecución de GA Math.	62
Tabla 6. Espacio de valores posibles y muestreados para los parámetros de MLP.	67
Tabla 7. Identificación, descripción y referencia para cada variable utilizada.	79
Tabla 8. Estadísticos básicos medidos sobre la población de modelos correctos. .	82
Tabla 9. Estadísticos básicos medidos sobre la población de modelos incorrectos.	83
Tabla 10. Ranking de variables.	84
Tabla 11. Selección de variables: algoritmo de Yu y Liu.	86
Tabla 12. Selección de variables: algoritmo de Koller y Sahami.	87
Tabla 13. Precisión y AUC de cada variable sobre los conjuntos de entrenamiento y prueba.	90
Tabla 14. Centroides estimados para las clases correcto e incorrecto.	91
Tabla 15. Precisión y AUC sobre los conjuntos de entrenamiento y prueba para los clasificadores GA Math..	92
Tabla 16. Precisión y AUC sobre los conjuntos de entrenamiento y prueba para cada clasificador obtenido con GA Logic.	94
Tabla 17. Valores optimizados para una y dos capas ocultas en la arquitectura del MLP.	97
Tabla 18. Optimización de algoritmos de búsqueda para Redes Bayesianas.	98
Tabla 19. Valores optimizados para cada tipo de kernel en SVM.....	99
Tabla 20. Resultados del Test de McNemar para la comparación de clasificadores univariados.	100
Tabla 21. Precisiones obtenidas por cada clasificador óptimo multivariable.	101
Tabla 22. Resultados del Test de McNemar para la comparación de clasificadores multivariados.	102
Tabla 23. Resultados del Test de DeLong para la comparación de clasificadores univariados.	104
Tabla 24. AUC obtenido para cada clasificador multivariable.	104
Tabla 25. Resultados del Test de DeLong para la comparación de clasificadores multivariados.	105
Tabla 26. Precisión de los mejores 10 clasificadores sobre los modelos incompletos del conjunto de prueba.....	107

RESUMEN

Uno de los problemas esenciales en la predicción computacional de la estructura tridimensional de proteínas corresponde a la evaluación de la calidad de un modelo proteico generado computacionalmente, esto es, clasificar cada modelo proteico en correcto o incorrecto. Este problema toma especial importancia cuando los modelos son generados por software automatizados a gran escala.

La mayoría de los métodos existentes para la evaluación de los modelos proteicos están basados en variables únicas que actúan como los clasificadores. La variable consistente en la energía libre total del sistema es aquella de mejor rendimiento cuando se le compara a otras variables o atributos del modelo proteico. Sin embargo, clasificadores multivariados basados en una serie de propiedades físicas, geométricas y estadísticas pueden mostrar un rendimiento significativamente mayor con respecto a los clasificadores de una variable, sobre todo para los casos más difíciles que corresponden a proteínas pequeñas y cuyo modelo obtenido computacionalmente es incompleto.

En el presente trabajo de tesis, se calcularon un total de 31 variables sobre un conjunto de modelos proteicos correctos e incorrectos generados con la técnica de modelado comparativo. Estas variables corresponden a propiedades del alineamiento secuencia-estructura entre la secuencia a modelar y la estructura molde, propiedades del modelo proteico generado, propiedades de la región del molde efectivamente utilizada para generar el modelo, y propiedades del molde completo utilizado para generar el modelo proteico. El conjunto de datos se dividió en conjuntos de entrenamiento, validación y de prueba. Se aplicaron distintos métodos de ranking, selección y extracción de variables para filtrar redundancia y maximizar la relevancia de las variables con respecto a la clase respuesta.

Luego, se aplicaron diferentes algoritmos de aprendizaje tales como redes bayesianas, máquinas de vectores de soporte, perceptrón multicapa y algoritmos genéticos con el fin de obtener clasificadores multivariados para el problema de la clasificación de un modelo en correcto e incorrecto. El rendimiento de cada clasificador multivariable, así como el rendimiento de cada variable única utilizada como clasificador, fue comparado con el rendimiento de los otros clasificadores con el fin de declarar a uno de ellos como aquel óptimo para el problema de la evaluación de la calidad de modelos proteicos generados computacionalmente.

El clasificador óptimo obtenido en este trabajo, generado con el algoritmo de aprendizaje de máquinas de vectores de soporte, presenta un aumento en el rendimiento de un 13% con respecto a los mejores clasificadores univariados.

1) Introducción

1.1) *Biología, el dogma central de la biología molecular, y la bioinformática*

La biología es el área de la ciencia que estudia la vida, en su forma más general. *Vida* es un término bastante amplio, y basta mirar alrededor de nosotros para encontrar diversas formas en las que esta se manifiesta. Dada esta tremenda diversidad, la biología se puede clasificar de acuerdo al tipo de vida que estudia: la botánica por ejemplo, se refiere al estudio de la vida de las plantas; la zoología estudia la vida de los animales, y la medicina estudia la salud humana.

En una primera instancia, estas ramas podrían considerarse independientes unas de otras. Sin embargo existen factores comunes a cada uno de los individuos que son objeto de estudio en cada una de ellas. Esta relación quedó en evidencia en la **teoría de la selección natural** planteada de manera paralela en el siglo XIX por Alfred Russel Wallace y Charles Darwin, dos destacados naturalistas británicos.

La teoría de la selección natural plantea que los seres vivos, sean éstos plantas, animales, etc., están sometidos a fuerzas evolutivas bajo las cuales ciertos fenotipos – características visibles – se ven favorecidos a lo largo del tiempo, de manera que los individuos adquieren gradualmente la capacidad de adaptarse al entorno que los rodea. De igual manera, los fenotipos que no son favorables para los individuos con respecto a su entorno, van haciéndose menos comunes con el paso del tiempo. Esta teoría está plasmada en *El Origen de las Especies*, de Charles Darwin [1, 2], publicada en 1859, y en ella se pueden encontrar innumerables ejemplos, generados a lo largo de más de veinte años de observaciones, en los que la selección natural se observa. Por ejemplo, Darwin observó que la forma y tamaño del pico de trece especies distintas de *pinzones* se adaptaba al tipo de comida que cada una de las especies consumía, argumentando que esa forma particular para cada especie se debe a la selección natural. (Fig 1)

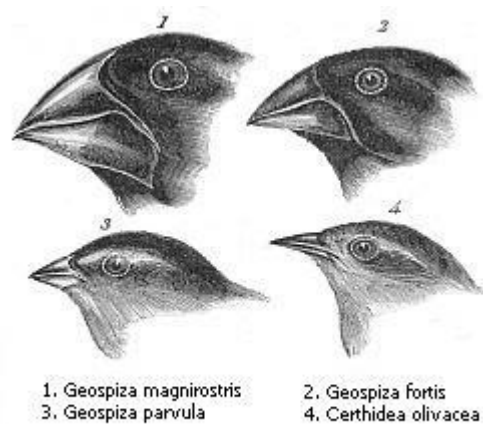


Fig 1. Los pinzones de Darwin. Ilustraciones de los picos de distintas especies de pinzones, adaptados de acuerdo a sus requerimientos de acceso a alimento.

Así como la selección natural se encarga de favorecer los fenotipos que permiten al individuo adaptarse a su entorno, también hay una fuente interna a cada organismo que determina los diversos fenotipos, y que corresponde al material genético llamado **genotipo**.

El material genético en los seres vivos está organizado en unidades llamadas cromosomas. La unidad básica de los cromosomas es el ácido desoxirribonucleico, o **ADN**, y a su vez la unidad básica del ADN es el ácido nucleico, del cual existen cuatro tipos: adenina (A), timina (T), citosina (C) y guanina (G). El ADN corresponde a moléculas de ácidos nucleicos dispuestas en forma de dos hebras helicoidales antiparalelas (Fig 2). La parte del material genético que **codifica** los fenotipos se conoce como genotipo. ¿Pero cómo ocurre esta decodificación de genotipo a fenotipo? Francis Crick, en su trabajo *On protein Synthesis* en 1958 enunció por primera vez el *Dogma Central de la Biología Molecular*, el cual explica este proceso.

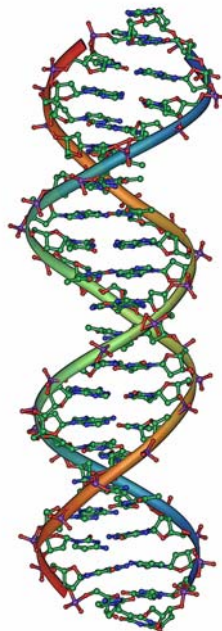


Fig 2. Molécula de ADN. El ADN tiene una estructura tridimensional y contiene la información que codifica el fenotipo del organismo.

El Dogma Central de la Biología Molecular (Fig 3) es la forma en que todo tipo de organismo vivo expresa su información genética, y consta de 3 etapas principales: *replicación, transcripción y traducción*.

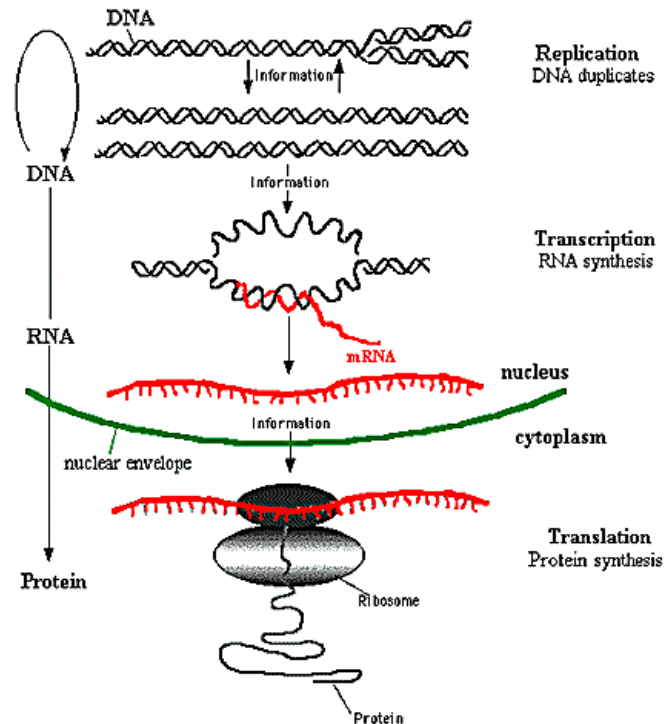


Fig 3. Dogma Central de la Biología Molecular. La replicación del ADN permite contar con el material genético. La transcripción genera ARN a partir del ADN. Una o más proteínas se generan a partir del ARN por un proceso de traducción. Estas proteínas serán las constituyentes básicas del fenotipo del organismo.

La **replicación** se refiere a la copia del ADN, lo que permite contar con el material genético necesario para generar los fenotipos. La **transcripción** corresponde al proceso por el cual se genera una o más hebras simples de ácido ribonucleico, o **ARN**, a partir del ADN. El ARN también está conformado por ácidos nucleicos pero con la diferencia de que la timina (T) se sustituye por el uracilo (U). La **traducción** corresponde al proceso de generación de una o más **proteínas** a partir del ARN. Las proteínas son polímeros compuestos por moléculas llamadas aminoácidos. Existen veinte aminoácidos diferentes, que se distinguen unos de otros de acuerdo a propiedades químicas.

La palabra proteína viene del griego “prota”, que significa “de principal importancia”. Y esta importancia es evidente, considerando que son las unidades básicas que conforman el fenotipo, y son las participantes claves de la mayoría de los procesos que ocurren en un organismo vivo, esto es, juegan un rol central en la estructura y función de los elementos que constituyen a un ser vivo. Así por ejemplo, el pico de las aves que Darwin estudiaba está compuesto principalmente por keratina, un tipo de proteína fibrosa con un alto contenido de un aminoácido llamado glicina. También los anticuerpos (o inmunoglobulinas), que protegen a muchos organismos de los ataques de virus y bacterias en el sistema inmune, son un tipo de proteína llamada glicoproteína y cuya estructura con forma de “Y” permite detectar los antígenos que estimulan la respuesta inmune.

En la última década se han desarrollado métodos experimentales que han permitido estudiar diferentes fenómenos biológicos desde un punto de vista global. Así, hoy es posible ver cómo los genes que componen el genoma de un organismo vivo se expresan en su conjunto (de acuerdo al dogma central de la biología molecular) frente a diferentes condiciones en el ambiente o a lo largo del tiempo. También es posible determinar el genoma completo de una especie determinada gracias a una técnica conocida como secuenciación. El uso de estas y otras técnicas experimentales ha generado en los últimos años enormes cantidades de información que se almacenan en bases de datos, y tanto el análisis de estos datos como el modelamiento de los fenómenos que los generan han aparecido como pasos naturales en el avance de las ciencias biológicas.

El uso, creación y desarrollo de herramientas computacionales, matemáticas y estadísticas para el análisis y modelamiento de fenómenos biológicos se conoce como **Bioinformática**.

1.2) Relación secuencia-estructura en proteínas. Relación estructura-función en proteínas.

Una proteína se puede modelar en su forma más simple como una secuencia de caracteres de largo variable en un alfabeto de tamaño veinte, correspondiente a cada uno de los aminoácidos. Por otro lado, existe un proceso en el cual las proteínas se conforman en estructuras tridimensionales complejas a partir de su secuencia primaria de aminoácidos (Fig 4). Este proceso ocurre principalmente debido a dos tipos de interacción: aquella entre las mismas moléculas que conforman a la proteína, y aquella entre las moléculas que conforman a la proteína y las moléculas del medio en el cual se encuentra. La estructura final, o conformación, que obtiene la proteína corresponde a aquella en la cual se alcanza el mínimo de energía libre accesible.

MET LEU SER ASP GLU ASP PHE LYS ALA VAL PHE GLY
MET THR ARG SER ALA PHE ALA ASN LEU PRO LEU TRP
LYS GLN GLN ASN LEU LYS LYS GLU LYS GLY LEU PHE



Fig 4. Plegamiento de una proteína. Proceso de conformación de la estructura nativa de una proteína a partir de su secuencia primaria de aminoácidos.

En los últimos cuarenta años se han reportado tres hallazgos experimentales independientes acerca de la relación secuencia/estructura en proteínas:

A principios de los años sesenta, Christian Anfinsen demostró que la información necesaria para que una proteína se pliegue a su conformación nativa en tres dimensiones bajo condiciones fisiológicas se encuentra completamente codificada en su secuencia primaria de aminoácidos [3, 4]. En particular, el experimento de Anfinsen mostró que si la estructura nativa de la enzima ribonucleasa (que cataliza la hidrólisis de RNA) era llevada a una forma lineal y después se le permitiera replegarse, entonces retomaría su forma nativa que le da su funcionalidad catalítica. El hecho de que la enzima tome su conformación nativa a pesar de la amplia gama de posibles conformaciones se explica porque la información contenida en la secuencia dicta un número de diferentes conformaciones posibles, y aquella más estable termodinámicamente es la conformación nativa, que es la funcionalmente activa. Este hallazgo sentó las bases para el desarrollo de las primeras técnicas de predicción computacional de la estructura tridimensional de proteínas a partir de la secuencia primaria de aminoácidos: la predicción *ab initio*.

A mediados de los ochenta, los científicos británicos Chothia y Lesk encontraron que la estructura es significativamente más conservada que la secuencia [5]. La principal conclusión de su trabajo es que la relación entre similitud de secuencias versus similitud estructural en proteínas no es lineal, sino que logarítmica (Fig 5). En otras palabras, proteínas que a nivel de secuencia comparten más de un 30-40% de identidad aminoacídica, tienen estructuras altamente similares. Este hallazgo llevó a postular que el número total de pliegues proteicos posibles sería un número limitado [6] y generó el nacimiento de dos nuevos métodos para la predicción computacional de la estructura tridimensional de proteínas, los cuales son ampliamente utilizados en la actualidad: el modelado por homología y el reconocimiento de pliegues, o plegamiento inverso.

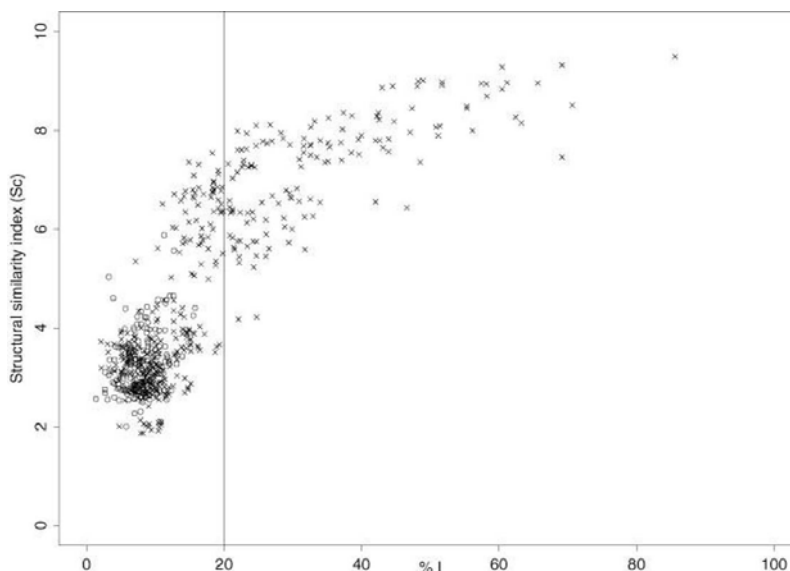


Fig 5. Relación secuencia – estructura en proteínas. La relación entre la similitud de secuencias (abscisa) y la similitud de estructuras (ordenada) en proteínas indica que un 40% de similitud de secuencia implica una alta similitud estructural.

En la década de los noventa, Holm y Sander demostraron que existen casos extremos donde secuencias proteicas totalmente diferentes (5-10% de identidad aminoacídica) y funcionalmente no relacionadas podían adoptar estructuras tridimensionales altamente similares [7-10].

En lo que se refiere a la relación secuencia/función y estructura/función en proteínas, debido a que será la disposición en tres dimensiones de los diferentes aminoácidos la que determinará la función puntual que la proteína llevará a cabo, resulta claro que la función proteica depende directamente y en mayor grado de la estructura tridimensional que de la secuencia primaria de aminoácidos; la alta sensibilidad y especificidad del sitio activo de una proteína está determinado principalmente por la conformación y entorno fisicoquímico específicos de los átomos que componen el sitio activo de la estructura nativa de la proteína.

1.3) La necesidad de la predicción computacional de la estructura tridimensional de proteínas.

En la actualidad existen más de tres millones de secuencias proteicas conocidas, lo cual contrasta fuertemente con las sólo 44,000 estructuras tridimensionales de proteínas que han sido resueltas experimentalmente [11, 12]. La principal razón de esta gran diferencia radica en el costo y en las dificultades técnicas envueltas en la determinación experimental para ambos casos, por lo cual lamentablemente esta diferencia seguirá creciendo de manera exponencial. La única solución para reducir esta brecha entre el número de secuencias y estructuras radica necesariamente en el desarrollo exitoso del proyecto mundial de genómica estructural [13] y además en la aplicación a gran escala de los métodos de predicción de estructuras tridimensionales de proteínas [14-18]

1.4) Métodos de predicción computacional de la estructura tridimensional de proteínas.

Actualmente existen tres técnicas diferentes para la predicción computacional de la estructura tridimensional de proteínas a partir de la secuencia aminoacídica: 1) modelado *ab initio* [19], 2) reconocimiento de pliegues o plegamiento inverso [20-23] y 3) modelado por homología [24, 25].

1.4.1) Modelado *ab initio*

El modelado *ab initio* consiste en la predicción de la estructura de una proteína usando exclusivamente su secuencia. El desarrollo de esta técnica es tremendamente importante, pero dado el alto grado de complejidad envuelto y el desconocimiento de cómo ocurre en detalle el proceso de plegamiento de las proteínas naturalmente, aún permanece como un método de predicción impreciso que genera muchas veces modelos alejados de la realidad.

1.4.2) Reconocimiento de pliegues o plegamiento inverso

Este método es denominado “plegamiento inverso” ya que consiste en buscar una o más estructuras conocidas dentro de las cuales una determinada secuencia se podría

eventualmente plegar. El reconocimiento de pliegues, tal como su nombre lo indica, predice más bien pliegues y no estructuras completas a un detalle atómico. Esta técnica consiste en producir múltiples modelos tridimensionales al reemplazar espacialmente los aminoácidos de la estructura conocida por aquellos contenidos en la secuencia que se quiere predecir (lo que se conoce como el alineamiento secuencia/estructura) y evaluar energéticamente dichos modelos. De acuerdo a esto, el pliegue es asignado o no a dicha secuencia.

1.4.3) Modelado comparativo o modelado por homología

El modelado comparativo predice la estructura tridimensional de una determinada secuencia proteica (que se denomina secuencia objetivo) basándose primariamente en su alineamiento entre la secuencia objetivo y estructura(s) molde(s), la construcción del modelo, y finalmente la evaluación del modelo construido (Fig 6).

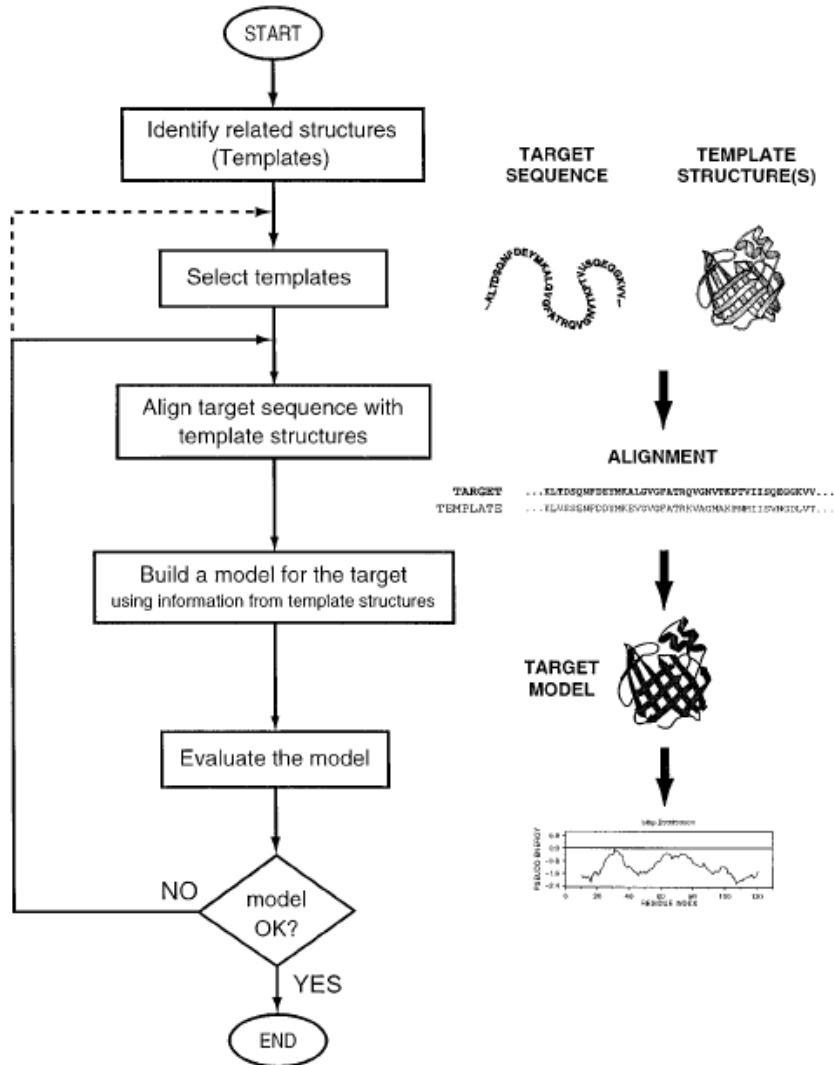


Fig 6. Modelado por homología. Este método de predicción de la estructura de una proteína a partir de su secuencia primaria de aminoácidos tiene como base la utilización de la información disponible para proteínas determinadas experimentalmente, para la generación del modelo proteico. Los pasos que se aplican son la identificación y selección de moldes apropiados para la secuencia a predecir, el aliniamiento secuencia-estructura, la generación del modelo proteico, y finalmente la evaluación del modelo generado. Figura adaptada de [25].

El modelado por homología comienza con la selección de homólogos con estructuras conocidas del repositorio Protein Data Bank, o PDB¹. La detección de homólogos se realiza comparando la secuencia objetivo con todas las secuencias de las estructuras presentes en PDB. Esto se logra usando métodos de alineamiento tales como el algoritmo de Needleman y Wunsch [26], BLAST o PSI-BLAST [27], entre otros. Si se tienen múltiples homólogos a la secuencia objetivo, el siguiente paso es seleccionar una o más estructuras que son las más apropiadas para construir el modelo. El criterio principal para definir cual será la estructura a utilizar para construir el modelo corresponde a la similitud de secuencia en el alineamiento: mayor similitud de secuencias está asociado a una mayor similitud estructural. Otros criterios para la determinación de las estructuras que se utilizan para la construcción del modelo involucran la presencia de inserciones o supresiones en el alineamiento, la pertenencia de la secuencia objetivo a una familia proteica específica, e incluso el ambiente fisicoquímico en que reside la estructura molde, en relación al ambiente de la proteína a modelar.

Una vez realizado el alineamiento entre la secuencia objetivo y la(s) estructura(s) molde(s), el siguiente paso corresponde a la construcción del modelo. Existen distintos métodos para la construcción del modelo que se distinguen unos de otros en cómo se utiliza la información presente en la estructura conocida. Los principales métodos utilizados corresponden a la construcción del modelo: (i) por unión de cuerpos rígidos [28] (o rigid body assembly), (ii) por calce de segmentos [29] (o segment matching) y (iii) por satisfacción de restricciones espaciales [30] (o satisfaction of spatial restraints). Este último es el método implementado en uno de los programas más utilizados para el modelado comparativo, MODELLER² y se basa en la generación de restricciones basadas en que la distancia entre un par de aminoácidos en el modelo es similar a la distancia de los correspondientes aminoácidos en la estructura molde. Esta restricción es complementada con otras restricciones estereoquímicas que se aplican sobre los ángulos de enlace, el largo de los enlaces, etc.

La última etapa del modelado comparativo corresponde a la evaluación del modelo generado, que permite decidir si este es adecuado o no. Existen dos categorías principales de funciones que se han propuesto para la evaluación del modelo: (i) funciones estadísticas de energía efectiva [31], que se basan en propiedades observadas de los aminoácidos que componen estructuras conocidas, y (ii) funciones físicas de energía efectiva [32], que se basan en la evaluación directa de la energía de solvatación de la proteína. Diferentes herramientas que permiten la evaluación del modelo generado se pueden clasificar en dos categorías: aquellas que chequean la correcta estereoquímica del modelo, y aquellas que verifican el ajuste de la secuencia a la estructura.

1.5) Errores posibles en el modelado comparativo de proteínas.

Cada vez que se predice la estructura tridimensional de una proteína se corre el riesgo de generar un modelo con distintos grados o niveles de error. La acumulación de errores conducirá a una pérdida gradual de la precisión del modelo, y por ende, a una pérdida de la capacidad para inferir correctamente la función de la proteína. Existen cinco tipos de

¹ <http://www.pdb.org>

² <http://salilab.org>

errores que se pueden producir cuando se aplica la técnica de modelado comparativo [25], los cuales se citan a continuación en orden creciente de importancia (Fig 7): a) modelado erróneo de la conformación de las cadenas laterales de aminoácidos, b) diferencias entre estructura molde y estructura objetivo en zonas correctamente alineadas, c) modelado erróneo de zonas sin estructura molde, d) alineamiento incorrecto entre estructura molde y secuencia objetivo, y e) selección errónea de estructura molde.

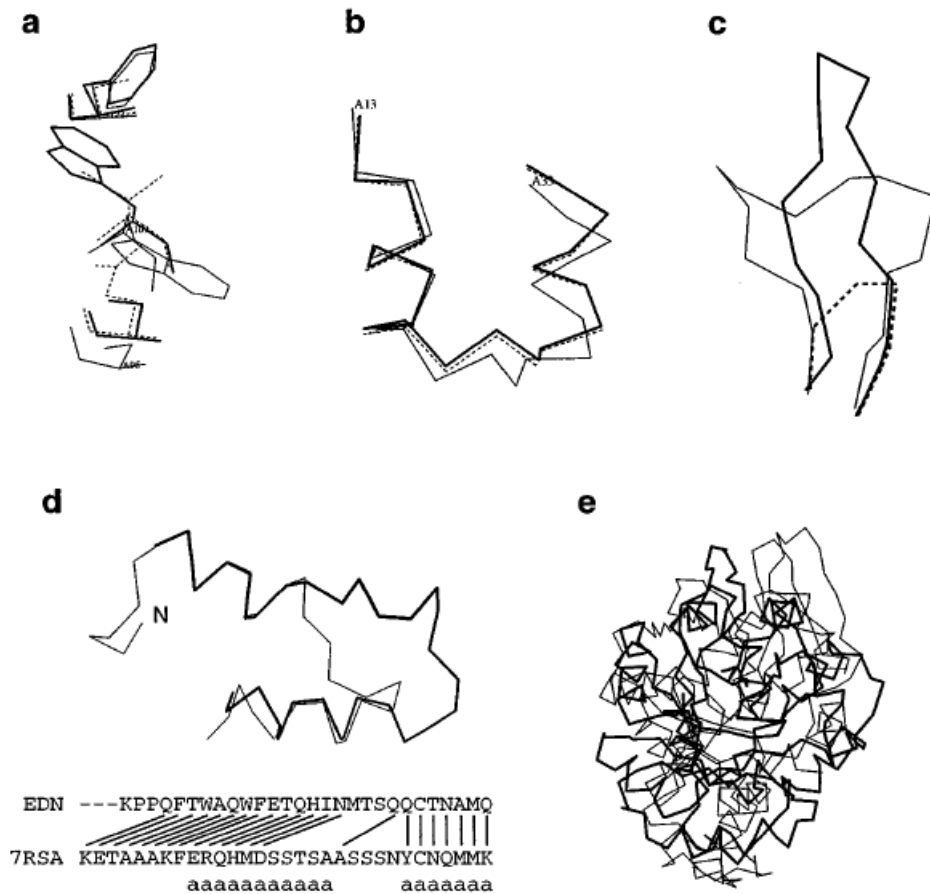


Fig 7. Errores en el modelado por homología. Los tipos de errores se muestran en orden creciente de importancia. a) Error en la conformación de las cadenas laterales. b) Diferencias entre estructura molde y estructura del modelo en zonas bien alineadas. c) Error en el modelado de zonas sin molde. d) Error en el alineamiento secuencia-estructura. e) Error en la selección del molde. Figura adaptada de [25].

1.6) Minería de Datos

Las técnicas más utilizadas para generar conocimiento a partir de una gran cantidad de datos o para agregar valor a la información disponible involucran principalmente redes neuronales [33], mapas auto organizativos [34, 35], algoritmos genéticos [36], clustering, análisis multivariado y análisis de componentes principales [37], árboles de decisión [38] y máquinas de vectores de soporte [39]. Estas técnicas tienen la bondad de explotar el hecho de que la mayoría de los procesos de toma de decisiones o clasificación que exhiben un buen rendimiento en problemas de la vida real, requieren de la combinación adecuada de múltiples variables o atributos que describen un determinado objeto a clasificar o categorizar.

En el caso de la evaluación de la calidad de los modelos tridimensionales de proteínas, la variable o atributo consistente en la energía libre total del sistema es aquella que exhibe un mejor rendimiento cuando se le compara a otras variables o atributos del modelo [40]. Sin embargo, cuando esta variable se combina de manera conveniente con otros atributos del modelo, la tasa de error en el proceso de clasificación disminuye considerablemente [16, 40, 41]. Existe un gran número de diversos atributos que pueden ser extraídos o calculados a partir de cada modelo, los cuales a su vez pueden generar una gran cantidad de particiones y/o combinaciones. Por lo tanto, sin importar el poder computacional de cálculo con el que se cuente, es impracticable evaluar todos los clasificadores posibles que se pueden generar. Debido a esto, las técnicas de reconocimiento de patrones y minería de datos mencionadas constituyen procedimientos bien definidos de manera que son capaces de generar soluciones cercanas al óptimo sin necesidad de realizar una búsqueda exhaustiva o completa.

Hasta la fecha, se han desarrollado dos trabajos que involucran herramientas de minería de datos y reconocimiento de patrones aplicadas al desarrollo de clasificadores óptimos de la calidad de estructuras tridimensionales de proteínas. El primero corresponde al desarrollo de un clasificador óptimo basado en redes neuronales para la evaluación de modelos generados mediante la técnica de reconocimiento de pliegues [16, 41]. El segundo corresponde al desarrollo de un clasificador óptimo basado en algoritmos genéticos para la evaluación de calidad de modelos tridimensionales de proteínas generados mediante la técnica de modelado comparativo [40]. En estos trabajos se ha demostrado el alto impacto que tiene en la disminución de la tasa de error de clasificadores óptimos la combinación adecuada de múltiples variables o atributos.

1.7) Objetivos

1.7.1) Objetivo General

El objetivo general de esta tesis es obtener un clasificador óptimo de la calidad de los modelos de estructuras tridimensionales de proteínas generados con la técnica de modelado comparativo a partir de la utilización de múltiples atributos correspondientes tanto a los modelos generados como a los moldes utilizados en su construcción. Este clasificador óptimo será integrado al módulo de evaluación del software de modelado comparativo MODELLER.

1.7.2) Objetivos Específicos

a) Seleccionar atributos no redundantes mediante medidas de asociación y dependencia tales como correlación clásica e información mutua, y técnicas de reducción de dimensionalidad como análisis de componentes principales. El propósito de esta primera etapa es sentar las bases para que los clasificadores a construir sean lo más simples posible. Así, con argumentos estadísticamente válidos se logrará descartar aquellos atributos o variables que aporten la menor información o mayor redundancia.

b) Obtener clasificadores basados en diversos algoritmos de aprendizaje.

c) Entrenar a cada clasificador obtenido con el conjunto de entrenamiento y en base a este aprendizaje clasificar los modelos pertenecientes al conjunto de prueba.

d) Comparar los resultados de los clasificadores obtenidos entre ellos y con aquellos ya existentes basados en algoritmos genéticos y árboles de decisión. El propósito de esta etapa es validar a uno del total de clasificadores como un clasificador óptimo dentro de aquellos que cuentan con procedimientos bien definidos y que ya han sido utilizados anteriormente en una serie de aplicaciones.

2) Marco Teórico

2.1) Generación de la base de datos.

Dada la directa relación entre estructura y función en proteínas, un paso que es clave en el modelado comparativo es aquél de **evaluación del modelo**, esto es, dado un modelo construido a partir de su secuencia aminoacídica, decidir si este es un modelo correcto o incorrecto.

Con el fin de mejorar el módulo de evaluación del modelo en el modelado comparativo, en el año 1,998 se procedió a modelar computacionalmente y de manera completa la base de datos de estructuras proteicas determinadas experimentalmente, Protein Data Bank, o PDB³.

El objetivo de modelar computacionalmente aquellas proteínas ya determinadas experimentalmente es determinar exactamente el grado de error que existe entre la estructura real de la proteína (asumiendo el error propio de la técnica experimental con la que se determinó) y la estructura predicha.

Esto implica que distintos criterios de qué es un modelo correcto e incorrecto, algunos más estrictos que otros, pueden aplicarse en términos del error en la superposición 3D de cada modelo con su estructura conocida. A su vez, esto permite poner a prueba distintos métodos de clasificación de los modelos al momento de ser evaluados, de manera de contar con una clasificación que minimice una función de costo que sea función de los falsos negativos, esto es, modelos correctos que se clasifican como incorrectos, y los falsos positivos, esto es, modelos incorrectos que se clasifican como correctos.

Así, un total de 10,553 modelos fueron construidos. Esta base de datos generada dio pie a una serie de estudios colaborativos entre el grupo de Bioinformática de la Pontificia Universidad Católica de Chile⁴, dirigido por Francisco Melo, y el grupo de Bioinformática de la Universidad de California en San Francisco⁵, dirigido por Andrej Sali que siguen hasta hoy con la realización de este trabajo, y que buscan la optimalidad de la etapa de evaluación de los modelos en el modelado comparativo.

Para este trabajo, la base de datos utilizada corresponde a un subconjunto de la base de datos inicial. El criterio con el que se determinó esta nueva base de datos involucra distintos parámetros que se explican a continuación:

a) Largo de la Cadena

Se refiere al número de aminoácidos que componen a la secuencia que está siendo modelada.

³ <http://www.pdb.org/>

⁴ <http://protein.bio.puc.cl/>

⁵ <http://salilab.org/>

b) Desviación Cuadrática Media (RMSD)

RMSD, o Root Mean Square Deviation, es una medida estadística típicamente utilizada para medir el error entre la estimación del valor de una variable u objeto y el verdadero valor que ésta tiene.

En términos estadísticos, si se tiene una variable θ y una estimación de esta variable, $\hat{\theta}$, entonces el RMSD está dado por

$$\sqrt{\mathbf{E}((\theta - \hat{\theta})^2)}$$

De igual manera, para una proteína se puede pensar que su verdadero valor corresponde a sus coordenadas en 3D, y su estimación corresponde a las coordenadas en 3D del modelo generado. Así, este índice mide el error de cada modelo generado con su correspondiente estructura determinada experimentalmente.

c) Porcentaje de carbonos alfa equivalentes.

Las proteínas son secuencias aminoacídicas, y cada aminoácido tiene dos extremos: uno que se conoce como el extremo amino, y otro que se conoce como el extremo carboxilo. Así, la unión de dos aminoácidos puede entenderse como la unión del extremo carboxilo del primer aminoácido con el extremo amino del segundo aminoácido. Cada aminoácido tiene además un carbono alfa (Fig 8). Así, es fácil imaginar una proteína como una secuencia de carbonos alfa que se distribuye en el espacio tridimensional.

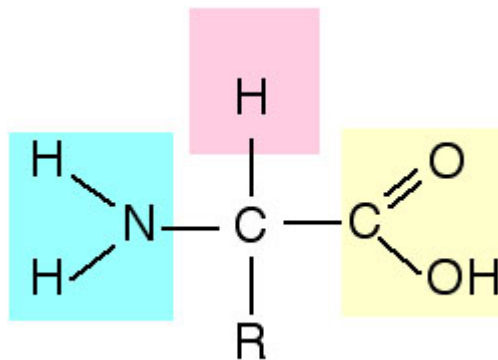


Fig 8. Estructura general de un aminoácido. El grupo izquierdo corresponde al grupo amino. El grupo derecho corresponde al grupo carboxilo. El carbono central corresponde al carbono alfa.

Si se considera que tanto la estructura real de la proteína como la estructura modelada computacionalmente se caracteriza por sus carbonos alfa, entonces se puede calcular el porcentaje de carbonos alfa equivalentes, que corresponde a la fracción de pares de carbonos alfa entre ambas estructuras que tienen una distancia menor o igual a 3.5 Angstroms (una vez que se han superpuesto óptimamente), y multiplicado por cien.

d) Porcentaje de identidad de secuencia en el alineamiento

Como se observa en Fig 6, una de las etapas del modelado comparativo corresponde al alineamiento de la secuencia a modelar con las proteínas relacionadas conocidas.

El porcentaje de identidad de secuencia en el alineamiento utilizado para construir el modelo se refiere al número de aminoácidos del modelo que calzan en posición con las estructuras presentes en PDB, dividido por el largo de la secuencia más corta, y multiplicado por cien.

Dada la definición de los parámetros utilizados para construir la base de datos sobre la que se trabaja, se procede a definir los modelos como correctos e incorrectos de la siguiente manera:

Se define como modelo correcto a aquél que cumpla con las siguientes condiciones:

- 1.- Largo de la cadena < 150 aminoácidos,
- 2.- RMSD global sobre carbono alfa < 7.0 Angstroms,
- 3.- Porcentaje de carbonos alfa equivalentes > 30%,
- 4.- Porcentaje de identidad de secuencia en el alineamiento utilizado para construir el modelo < 40%.

Se define como modelo incorrecto a aquél que cumpla con las siguientes condiciones:

- 1.- Largo de la cadena < 150 aminoácidos,
- 2.- RMSD sobre carbono alfa > 9.0 Angstroms,
- 3.- Porcentaje de carbonos alfa equivalentes < 15%,
- 4.- Porcentaje de identidad de secuencia en el alineamiento utilizado para construir el modelo < 90% y > 14%.

La aplicación de estas definiciones sobre la base de datos original de 10,553 modelos genera una nueva base de datos con 2,299 modelos, 1,153 modelos correctos - marcados como clase 1 - y 1,146 modelos incorrectos, marcados como clase -1 (Fig 9).

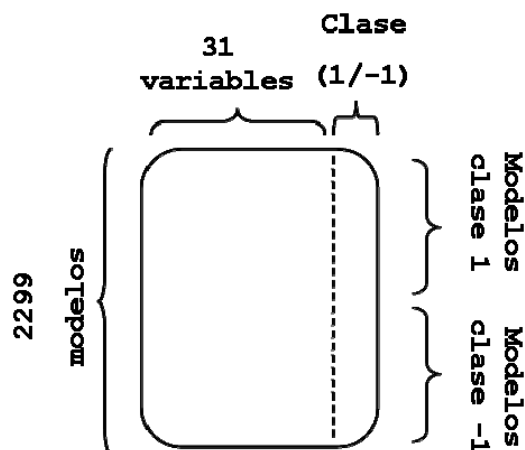


Fig 9. Esquema de la base de datos. La base de datos generada cuenta con 2299 modelos, 1153 de ellos correctos y 1146 incorrectos. Para cada modelo se midieron 31 propiedades que involucran el alineamiento secuencia-estructura, la estructura del modelo generado, la estructura efectiva del molde utilizado y la estructura completa del molde (ver sección 2.2 para más detalles). Cada modelo correcto se asoció con el valor “1”, y cada modelo incorrecto con el valor “-1”.

Es importante señalar que los modelos correctos no son modelos realmente buenos en términos de su precisión con respecto a la estructura conocida. El criterio de que el largo de la cadena sea menor a 150 en ambas clases se debe a que la discriminación sobre el espacio de modelos pequeños es más difícil que aquella sobre el espacio de modelos grandes. El clasificador óptimo a obtener en este trabajo será incorporado en el módulo de evaluación del software de modelado comparativo MODELLER para evaluar la calidad de aquellos modelos que ya fueron previamente clasificados como incorrectos de acuerdo a una de las etapas del esquema de clasificación del módulo de evaluación. Así, cada modelo que llegue a ser evaluado por el clasificador óptimo a obtener en este trabajo sólo podrá clasificarse en una de dos categorías: modelo medianamente correcto, que son aquellos de mediana calidad, y modelo definitivamente incorrecto, que son modelos de mala calidad. Para efectos de este trabajo, hablaremos directamente de modelo correcto e incorrecto.

En trabajos anteriores [42] se han generado otras bases de datos en los cuales los modelos correctos son muy precisos y los incorrectos son muy imprecisos, lográndose una clasificación con errores muy bajos en términos de la tasa de falsos positivos y falsos negativos.

2.2) Descripción de las variables medidas sobre los modelos.

Las variables medidas sobre los modelos se pueden clasificar de acuerdo a cuatro categorías principales:

I.- Variables medidas sobre el alineamiento secuencia-estructura.

Estas variables permiten capturar información relevante a partir del alineamiento entre la secuencia objetivo (o target sequence) y la estructura molde (o template) en el modelado comparativo (Fig 10). Las variables medidas fueron:

- 1) Porcentaje de identidad de secuencia en el alineamiento.
- 2) Z-score del alineamiento.

- 3) E-value de PSI-BLAST.
- 4) Fracción de la cadena que fue posible modelar.

II.- Variables medidas sobre la estructura de cada modelo.

Corresponden a variables estadísticas, geométricas y físicas medidas a partir de las estructuras de los modelos generados (Fig 10).

III.- Variables medidas sobre la estructura parcial del molde utilizado para construir el modelo.

Corresponden a variables estadísticas, geométricas y físicas medidas a partir de aquella región de la estructura molde (o template) que efectivamente se utilizó para modelar la correspondiente región de la secuencia objetivo (Fig 10).

IV.- Variables medidas sobre la estructura completa del molde utilizado para construir el modelo.

Corresponden a variables estadísticas, geométricas y físicas medidas a partir de la estructura molde completa, sin importar qué proporción de ella sirvió efectivamente para modelar la secuencia objetivo (Fig 10).

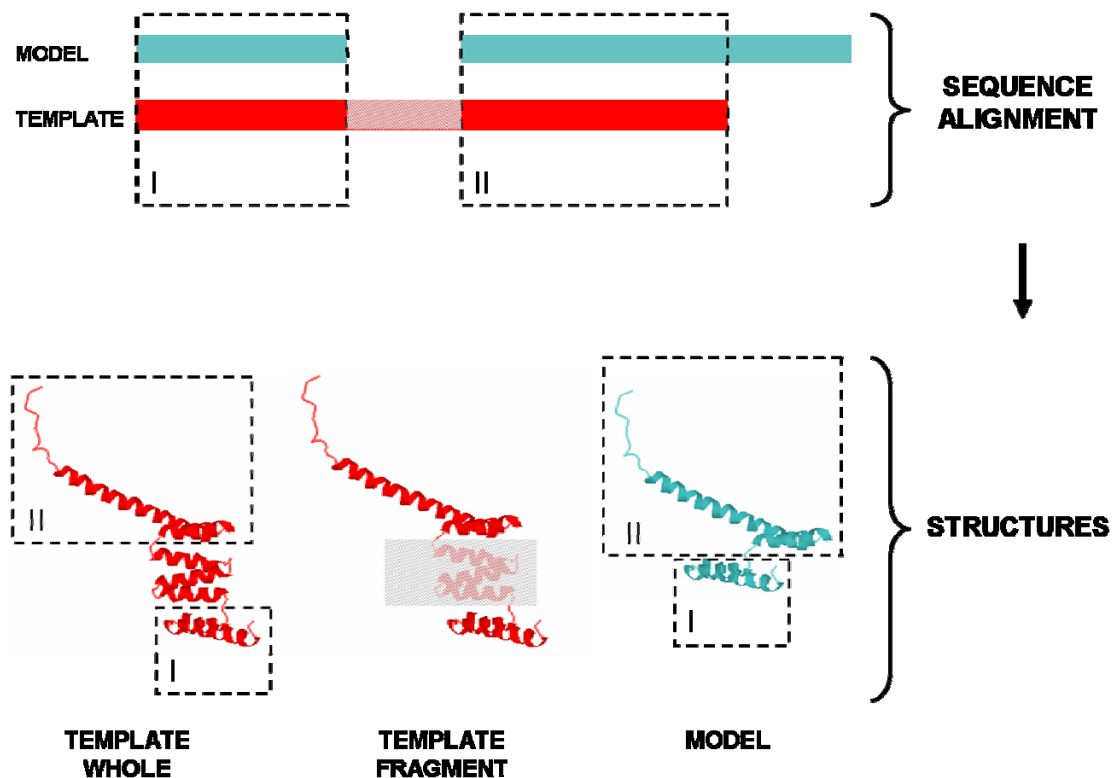


Fig 10. Categorías utilizadas en la medición de variables. Las variables medidas sobre los modelos corresponden a propiedades del alineamiento secuencia-estructura, propiedades del modelo proteico generado, propiedades del fragmento del molde utilizado para generar el modelo, y propiedades del molde completo utilizado para generar el modelo.

Las variables medidas en las categorías II, III y IV fueron:

- 5) Largo de la cadena.
- 6) Compactación.
- 7) Z-score de energía combinada.
- 8) Z-score de energía para pares residuo-residuo.
- 9) Z-score de energía de la superficie accesible.
- 10) Propensión a la partición.
- 11) Orden de contacto absoluto.
- 12) Orden de contacto relativo.
- 13) Radio de giro.

El concepto de **potencial estadístico** juega un papel clave en la definición de aquellas variables que involucran un Z-score, por lo que se explica a continuación.

Los potenciales estadísticos (también conocidos como knowledge-based potentials) constituyen la herramienta más utilizada para evaluar la calidad de los modelos generados mediante la aplicación de técnicas computacionales de predicción de la estructura tridimensional de proteínas. Se derivan a partir de datos experimentales y tienen la virtud de ser capaces de extraer de manera simple una gran porción de la complejidad y diversidad de las interacciones moleculares existentes en la conformación nativa (la estructura real) de las proteínas. Por esta razón, los potenciales estadísticos constituyen el esqueleto central de la gran mayoría de los métodos de predicción computacional de estructuras tridimensionales de proteínas [14, 16, 17] como también de los métodos de evaluación de la calidad de los modelos generados [23, 42-49].

Los potenciales estadísticos se calculan mediante la aplicación de la ley inversa de Boltzmann [31, 50]. La ley de Boltzmann relaciona la probabilidad p de un estado descrito por las variables i,j,k con la energía E_{ijk} de dicho estado, respecto al total de conformaciones posibles del sistema:

$$P_{ijk} = \frac{e^{\frac{-E_{ijk}}{k_B T}}}{\sum_{ijk} e^{\frac{-E_{ijk}}{k_B T}}}$$

donde k_B es la constante de Boltzmann y T es la temperatura del sistema medida en

Kelvin. Si llamamos $C = \sum_{ijk} e^{\frac{-E_{ijk}}{k_B T}}$, entonces la ley inversa de Boltzmann está dada por:

$$E_{ijk} = -k_B T \ln(P_{ijk}) - k_B T \ln(C)$$

Si consideramos a i,j como un par de átomos y k como la distancia entre ellos, en una primera etapa se pueden obtener las densidades observadas en cada estado i,j,k utilizando para esto las estructuras determinadas experimentalmente. Luego, estas densidades se utilizan como estimaciones de la probabilidad y son transformadas a energía junto a una serie de correcciones estadísticas [51].

Así, se obtiene un gran conjunto de funciones de energía versus distancia para cada par de átomos, según lo observado en los datos experimentales. Aquellas interacciones observadas con alta frecuencia en las proteínas nativas tendrán asociada una energía favorable, que corresponde a un valor negativo. Por el contrario, aquellas interacciones observadas con baja frecuencia en las estructuras nativas tendrán asociada una energía desfavorable, que corresponde a un valor positivo. Por último, aquellas interacciones que no se ven favorecidas ni desfavorecidas en las estructuras nativas, tendrán una energía nula.

Dado que cada estructura tiene sus propias redes de interacción y una composición y número variable de aminoácidos, los valores absolutos de energía no tienen ningún sentido, por lo cual se deben utilizar distribuciones de energía que sean puntos de referencia para evaluar el valor estadístico de la energía obtenida en cada caso.

Para esto se utilizan los Z-scores de energía, los cuales se encuentran normalizados en base a secuencias aminoacídicas aleatorias de idéntica composición y conformación tridimensional [42, 50]. Esto proporciona un marco de referencia adecuado, de manera de poder comparar directamente los valores calculados para diferentes estructuras sin importar su pliegue, tamaño y composición aminoacídica.

El **Z-score de energía de una estructura** viene dado por:

$$Z_e = \frac{E_e - \mu_A}{\sigma_A}$$

Donde Z_e corresponde al Z-score de energía de la estructura, E_e corresponde a la energía absoluta total de la estructura, y μ_A y σ_A corresponden a la media y desviación estandar de los valores de energía de las secuencias aleatorizadas, respectivamente.

El concepto de potencial estadístico nos permite definir muchas de las variables medidas sobre las cuatro categorías principales. A continuación se definen las variables de las categorías I, II, III y IV.

- 1) Porcentaje de identidad de secuencia en el alineamiento: Es el número total de residuos idénticos en el alineamiento secuencia-estructura, dividido por el largo de la secuencia más corta en el alineamiento, y multiplicado por 100 (Fig 11A).
- 2) Z-score del alineamiento: El Z-score del alineamiento secuencia-estructura corresponde al valor $\frac{\mu - S}{\sigma}$, donde S es la suma de los puntajes obtenidos de la matriz de similitud de residuos⁶ del software MODELLER utilizado para el modelado comparativo [30]. Esta matriz mide para cada par de residuos un puntaje (o score) de acuerdo a diferentes propiedades observadas para ese par en distintas familias de proteínas (Fig 11B). El valor μ es el promedio de la distribución de 200 puntajes obtenidos intercambiando los pares de residuo en

⁶ Un residuo es un término general para la unidad que compone un polímero. En el contexto de las proteínas, corresponde a un aminoácido.

cada una de las secuencias a modelar y del molde, y σ es la desviación estandar de esta distribución.

- 3) E-value de PSI-BLAST: BLAST – Basic Local Alignment Search Tool – es una herramienta ampliamente utilizada en el área de la bioinformática para encontrar regiones de similitud local entre secuencias. El programa compara secuencias proteicas con secuencias disponibles en bases de datos, y calcula la significancia estadística de las identidades de secuencia obtenidas. PSI-BLAST – Position Specific Iterative BLAST – es una variante de BLAST que permite encontrar iterativamente proteínas cada vez más distantes pero relacionadas para una secuencia proteica dada. El E-value (Expected value) de PSI-BLAST [27] es un parámetro que describe el número de calces que se espera observar al azar cuando se genera el alineamiento sobre una base de datos de un tamaño determinado.
- 4) Fracción modelada (target coverage): Es la fracción de la secuencia objetivo, que fue efectivamente modelada (Fig 11C).
- 5) Largo de la cadena: Para la categoría II, corresponde al número de aminoácidos en la secuencia que fue posible modelar. Para la categoría III, corresponde al número de aminoácidos en la secuencia proteica utilizada como molde que efectivamente se utilizó para modelar. Para la categoría IV, corresponde al número de aminoácidos en la secuencia proteica completa utilizada como molde, sin importar qué proporción de ella sirvió efectivamente para modelar la secuencia objetivo.

- 6) Compactación: Corresponde a $\frac{\sum_{i=1}^n v_i^{AA}}{V}$, donde n es el número de residuos en la proteína, v_i^{AA} es el volumen del residuo i y V es el volumen de la esfera [52] cuyo diámetro está dado por el par de átomos más lejano en la proteína.

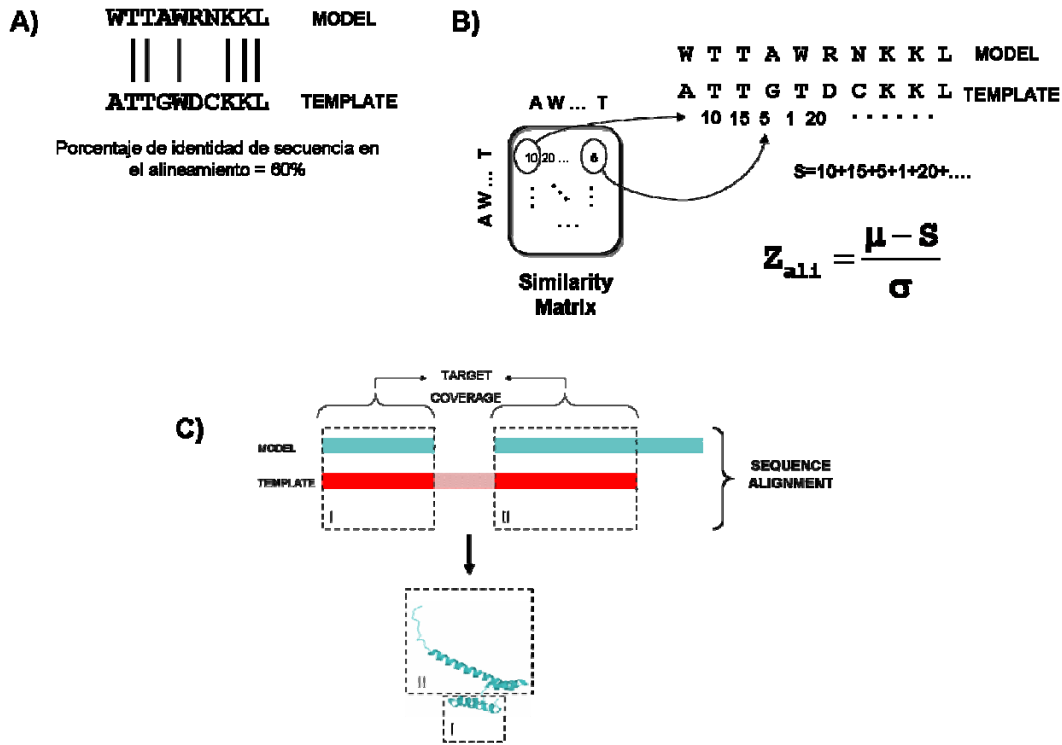


Fig 11. Propiedades del alineamiento secuencia–estructura. A) Porcentaje de identidad del alineamiento, B) Z-score del alineamiento, calculado a partir de los valores de la matriz de similitud, y C) Fracción modelada, que corresponde a la fracción de la secuencia objetivo que fue posible modelar.

7) Z-score de energía para pares residuo-residuo: Los potenciales estadísticos para pares residuo-residuo se calculan utilizando la siguiente ecuación [42]:

$$E_k^{ij}(\mathbf{l}) = RT \ln(1 + M_{ijk} \sigma) - RT \ln \left[1 + M_{ijk} \sigma \frac{f_k^{ij}(\mathbf{l})}{f_k^{xx}(\mathbf{l})} \right]$$

donde M_{ijk} es el número de ocurrencias para el par de residuos (i, j) separados por k residuos en la secuencia, esto es, $k = |I - J|$, donde I y J son los índices de los residuos (i, j) , respectivamente:

$$M_{ijk} = \sum_{l=1}^n f(i, j, k, l)$$

donde n es el número de clases de distancia (número de bins), $\sigma = \frac{1}{50}$ es el peso dado a cada observación, tal que con 50 observaciones $f_k^{ij}(\mathbf{l})$ y $f_k^{xx}(\mathbf{l})$ tendrán pesos iguales para el cálculo de $E_k^{ij}(\mathbf{l})$. $f_k^{ij}(\mathbf{l})$ es la frecuencia relativa de ocurrencias del par (i, j) con separación de secuencia k en la clase de distancia \mathbf{l} .

$$f_k^{ij}(\mathbf{l}) = \frac{f(\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l})}{M_{i,jk}}$$

$f_k^{xx}(\mathbf{l})$ es la frecuencia de ocurrencia relativa sobre todos los pares (\mathbf{i}, \mathbf{j}) a separación de secuencia \mathbf{k} en la clase de distancia \mathbf{l} :

$$f_k^{xx}(\mathbf{l}) = \frac{\sum_{i=1}^r \sum_{j=1}^r f(\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l})}{\sum_{i=1}^r \sum_{j=1}^r \sum_{k=1}^m f(\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l})}$$

en que r es el número de pares (\mathbf{i}, \mathbf{j}) y m es el número de clases para la separación de secuencia. La temperatura T se fija en 293 K, por lo que $RT = 0.582 \frac{\text{kcal}}{\text{mol}}$, donde R es la constante de los gases ideales.

Luego, el Z-score se calcula utilizando la ecuación $z_k = \frac{E_k - \mu_A}{\sigma_A}$.

8) Z-score de energía de la superficie accesible: La superficie accesible de un átomo se define como el número de átomos que se encuentran dentro de una esfera, alrededor del átomo central; el radio de la esfera es el rango de distancia del potencial estadístico. El potencial estadístico se calcula como sigue [42]:

$$E^i(\mathbf{r}) = RT \ln(1 + M_i \sigma) - RT \ln\left(1 + M_i \sigma \frac{f^i(\mathbf{r})}{f_{\text{ref}}^i(\mathbf{r})}\right)$$

M_i es la frecuencia del átomo central \mathbf{i} sobre todas las clases que componen a la esfera:

$$M_i = \sum_{r=1}^R f(\mathbf{i}, \mathbf{r})$$

donde R es el número de clases que componen a la esfera y $\sigma = \frac{1}{50}$ representa al peso dado a cada observación. $f^i(\mathbf{r})$ es la frecuencia relativa de ocurrencia del átomo central \mathbf{i} en la clase \mathbf{r} :

$$f^i(\mathbf{r}) = \frac{f(\mathbf{i}, \mathbf{r})}{M_i}$$

y $f_{\text{ref}}^i(\mathbf{r})$ corresponde a la frecuencia sobre un átomo central que no tiene preferencia por ninguno de los estados accesibles:

$$f_{\text{ref}}^i(\mathbf{r}) = \frac{M_i}{R} = \frac{1}{R}$$

Luego, el Z-score se calcula utilizando la ecuación $z_r = \frac{E_r - \mu_A}{\sigma_A}$.

- 9) Z-score de energía combinada: se define como la suma de las energías normalizadas de pares residuo-residuo y la energía de superficie accesible. La normalización se logra dividiendo cada tipo de energía por su desviación estandar.

Luego, el Z-score se calcula utilizando la ecuación $z_e = \frac{E_e - \mu_A}{\sigma_A}$.

- 10) Propensión a la partición: es una medida de cuán efectivamente se encuentra enterrada la superficie hidrofóbica en una estructura dada. [53]. Se define como:

$$\pi = \frac{2n_c}{q_H n_H}$$

donde n_c es el número de contactos en la estructura, q_H es el número de coordinación promedio de un residuo hidrofóbico (el número de coordinación se define como el número de vecinos que tiene el átomo central) y n_H es el número de residuos hidrofóbicos en la proteína. La interpretación física del término $2n_c$ es el número de sitios de coordinación involucrados en todos los contactos residuo-residuo, y es proporcional a la superficie enterrada de la proteína. $q_H n_H$ es el total de sitios de coordinación hidrofóbicos, y es proporcional a la superficie total hidrofóbica.

- 11) Orden de contacto absoluto: Corresponde a la separación promedio **en la secuencia** de residuos que se encuentran a una distancia menor o igual a 8 Angstroms **en la estructura** [54] (Fig 12A).
- 12) Orden de contacto relativo: Corresponde al orden de contacto absoluto dividido por el largo de la cadena (Fig 12A).
- 13) Radio de giro: Es la raíz cuadrada de la distancia atómica promedio entre los centroides de las cadenas laterales de los residuos y el centro de masa de la estructura (Fig 12B).

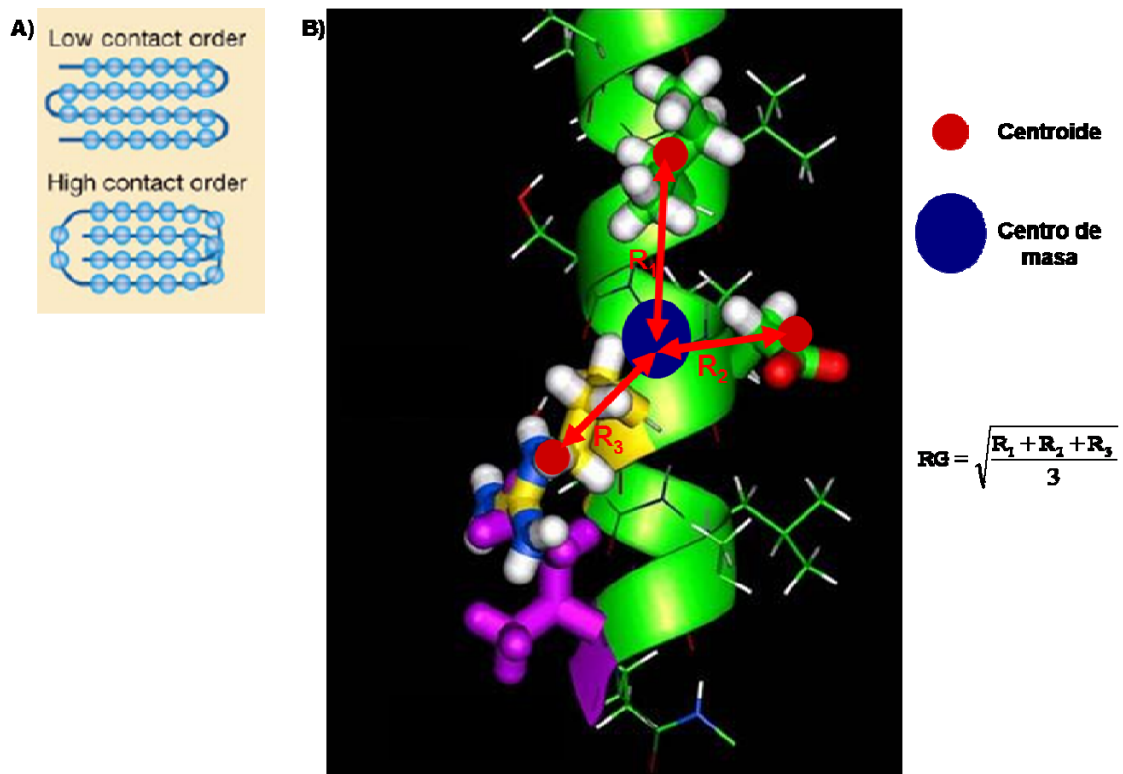


Fig 12. Propiedades estructurales medidas sobre los modelos. A) El orden de contacto absoluto mide la separación en la secuencia de aquellos aminoácidos que se encuentran a una distancia menor o igual a 8 Angstroms en la estructura. El orden de contacto relativo corresponde al orden de contacto absoluto con respecto al largo de la cadena. Figura adaptada de [55]. B) El radio de giro mide la distancia promedio de los centroides al centro de gravedad de la estructura.

2.3) Estadística descriptiva de los datos y detección de outliers.

Tanto para la población de modelos correctos como incorrectos, se puede representar una variable por \mathbf{x}_{ij} , donde i corresponde al modelo i -ésimo y j corresponde a la variable j -ésima. Para cada variable j , se calcularon distintos estadísticos en ambas poblaciones (Tabla 1).

Tabla 1. Definición de estadísticos básicos.

Estadístico	Fórmula
Mínimo	$\text{MIN}\{x_{ij}\}_i$
Máximo	$\text{MAX}\{x_{ij}\}_i$
Rango	$\text{MAX}\{x_{ij}\}_i - \text{MIN}\{x_{ij}\}_i$
Promedio	$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$
Varianza	$s_j^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$
Desviación Estandar	$s_j = \sqrt{s_j^2}$
Skewness	$\frac{\sqrt{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^3}{\left(\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \right)^{\frac{3}{2}}}$
Kurtosis	$\frac{n \sum_{i=1}^n (x_{ij} - \bar{x}_j)^4}{\left(\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \right)^2} - 3$

Skewness: medida de la asimetría de la distribución de cada variable j . Kurtosis: medida del grado de concentración en la zona central de la distribución de cada variable j .

Las medidas de skewness y kurtosis se utilizaron como indicadores de la normalidad de los datos. Una distribución perfectamente normal debe tener tanto un skewness como kurtosis iguales a cero. Adicionalmente, una herramienta gráfica que permitió visualizar la normalidad de los datos corresponde a los gráficos de probabilidad normal (normal probability plots [56]). En estos gráficos, los valores que los modelos toman para cada variable son graficados contra una distribución normal teórica, de manera que los puntos debieran formar una línea recta. El hecho de que los puntos se alejen de esta línea recta es un indicador claro de que los datos no se distribuyen normalmente.

La normalidad de las variables es algo muy útil de verificar, pues la mayoría de las herramientas de inferencia estadística, tales como intervalos de confianza y test de hipótesis, son paramétricos normales, esto es, al ser aplicados **asumen** que los datos se distribuyen normalmente.

La detección de outliers corresponde a la búsqueda de valores sobre las variables con error experimental, los cuales se generaron al momento de construir la base de datos o por otra causa.

La primera herramienta que se debe utilizar para la detección de outliers corresponde al conocimiento del dominio de cada variable. Luego, la herramienta gráfica más utilizada para la detección de outliers corresponde a los gráficos de caja y bigotes, también llamada box-and-whisker plots o simplemente boxplots, propuesta por el estadístico americano John Tukey en 1977.

Cada Boxplot representa una división por cuartiles del rango de valores que toma cada variable, esto es, los datos se ordenan y se dividen en cuatro grupos de manera tal que:

- el primer cuartil (Q_1) es el valor que corta los datos en el 25% inferior de valores.
- el segundo cuartil (Q_2), también llamado mediana, es el valor que corta los datos en el 50%.
- El tercer cuartil (Q_3), es el valor que corta los datos en el 25% superior de valores.

Se define el Rango Intercuartil, o Interquartile Range, como $IQR=Q_3-Q_1$ y utilizando estos valores, se define como un outlier a todo punto que cumpla con cualquiera de las siguientes condiciones:

$$\begin{aligned} X_{ij} &< Q_1 - 1.5 \cdot IQR \\ X_{ij} &> Q_3 + 1.5 \cdot IQR \end{aligned}$$

A su vez, se define como un outlier lejano (o far outlier) a todo aquél punto que cumpla con cualquiera de las siguientes condiciones:

$$\begin{aligned} X_{ij} &< Q_1 - 3.0 \cdot IQR \\ X_{ij} &> Q_3 + 3.0 \cdot IQR \end{aligned}$$

Es importante señalar que el criterio utilizado por Tukey para definir los valores 1.5 y 3.0 corresponde a asumir que las variables siguen una distribución normal, por lo que el no cumplimiento de la normalidad en las variables invalida en gran medida este método.

A continuación se presentan los métodos utilizados para el ranking, selección de variables y extracción de nuevas variables a partir de las originales. Los métodos de ranking establecen un orden de las variables de acuerdo a la relevancia que tienen para la clase respuesta. Los métodos de selección de variables (o feature selection) definen un subconjunto de las 31 variables iniciales que serán las más relevantes para la clase respuesta y las menos redundantes entre ellas. Los métodos de extracción de variables (o feature extraction) evidencian la redundancia presente de acuerdo a la combinación de las variables originales para obtener nuevas variables.

Cabe notar que el número de variables utilizadas no representa desafío alguno en términos computacionales para los métodos de aprendizaje que generan posteriormente los clasificadores, por lo que podrían evitarse las etapas de ranking, selección y extracción de variables. Sin embargo, uno de los objetivos es generar entendimiento a partir de los datos que se están utilizando, sobre todo por el significado biológico, físico, y geométrico que pueda tener cada una de las variables. Así, es de sumo interés contar con herramientas que permitan entender cómo cada variable es relevante para la clase respuesta, y cómo estas variables son redundantes en términos de la información que proporcionan.

2.4) Métodos de ranking de variables.

Con el fin de conocer cuan relevante es cada variable para la clase respuesta en términos de su poder de discriminación entre modelos correctos e incorrectos, es necesario disponer de métricas que permitan determinar el grado de asociación entre cada variable y la clase respuesta. Para este fin, se utilizaron medidas basadas en estadística clásica y teoría de la información.

2.4.1) Método de ranking basado en estadística clásica

El índice más utilizado para medir el grado de asociación entre dos variables \mathbf{X} e \mathbf{Y} es el coeficiente de correlación lineal

$$r_{\mathbf{x},\mathbf{y}} = \frac{\sum_{i=1}^n (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)}{\sqrt{\sum_{i=1}^n (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)^2} \cdot \sqrt{\sum_{i=1}^n (\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)^2}}$$

con:

$r_{\mathbf{x},\mathbf{y}} = 1$ indicando una relación estrictamente lineal con pendiente positiva entre las variables,

$0 < r_{\mathbf{x},\mathbf{y}} < 1$ indicando una tendencia lineal positiva entre las variables,

$r_{\mathbf{x},\mathbf{y}} = 0$ indicando la no existencia de asociación entre las variables,

$0 > r_{\mathbf{x},\mathbf{y}} > -1$ indicando una tendencia lineal negativa entre las variables, y

$r_{\mathbf{x},\mathbf{y}} = -1$ indicando una relación estrictamente lineal con pendiente negativa entre las variables.

Lamentablemente este índice de asociación entre variables sólo puede ser utilizado cuando ambas variables son cuantitativas.

Para el caso en que se quiere medir la asociación entre una variable cuantitativa y una variable cualitativa, se recurre al **grado de relación funcional** entre ambas variables. Sea \mathbf{X} una variable cualitativa con q categorías o modalidades, e \mathbf{Y} una variable cuantitativa. Sean

$$\mathbf{Y}_{1j}, \mathbf{Y}_{2j}, \dots, \mathbf{Y}_{n_jj}$$

las observaciones que toma la variable \mathbf{Y} sobre la modalidad j de \mathbf{X} , tal que $\sum_{j=1}^q n_j = \mathbf{n}$.

Así, el grado de relación funcional está dado por

$$\eta_{\mathbf{y}|\mathbf{x}}^2 = \frac{\mathbf{b}^2}{\mathbf{s}_y^2}$$

con

$$\mathbf{b}^2 = \sum_{j=1}^q \frac{\mathbf{n}_j}{\mathbf{n}} (\bar{\mathbf{y}}_j - \bar{\mathbf{y}})^2 \text{ la variabilidad total entre los grupos (correcto e incorrecto),}$$

$$\mathbf{w}_j^2 = \frac{1}{\mathbf{n}_j} \sum_{i=1}^{\mathbf{n}_j} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)^2 \text{ la variabilidad dentro de cada grupo } j,$$

$$\mathbf{w}^2 = \sum_{j=1}^q \frac{\mathbf{n}_j}{\mathbf{n}} \mathbf{w}_j^2 \text{ la media ponderada de la variabilidad intra-grupo,}$$

$$\mathbf{s}_Y^2 = \frac{1}{\mathbf{n}} \sum_{j=1}^q \sum_{i=1}^{\mathbf{n}_j} (\mathbf{y}_{ij} - \bar{\mathbf{y}})^2 \text{ la variabilidad total de la variable } Y,$$

$$\bar{\mathbf{y}}_j = \frac{1}{\mathbf{n}_j} \sum_{i=1}^{\mathbf{n}_j} \mathbf{y}_{ij} \text{ el promedio del grupo } j,$$

$$\bar{\mathbf{y}} = \frac{1}{\mathbf{n}} \sum_{j=1}^q \bar{\mathbf{y}}_j \text{ el promedio total.}$$

Es posible demostrar que la variabilidad total de las observaciones se descompone en la variabilidad intragrupo y la variabilidad intergrupo de la siguiente manera:

$$\mathbf{s}_Y^2 = \mathbf{w}^2 + \mathbf{b}^2$$

Así, el grado de relación funcional toma valores entre 0 y 1.

Si $\eta_{Y|X}^2 = 1$ hay relación funcional estricta.

Si $\eta_{Y|X}^2 = 0$ no hay relación entre \mathbf{X} e \mathbf{Y} .

En el contexto de este trabajo, la variable \mathbf{X} corresponde a la clase respuesta, esto es, modelo correcto (clase 1) o incorrecto (clase -1), y dado que es binaria entonces $q=2$. La variable \mathbf{Y} corresponde a cualquiera de las 31 variables medidas sobre los modelos.

2.4.2) Métodos de ranking basados en teoría de la información

La teoría de la información es una rama de la matemática que nace de la necesidad de transmitir información a través de un canal ruidoso. Su desarrollo se atribuye principalmente al trabajo de Claude Shannon [57], quien en 1948 definió matemáticamente el concepto de entropía.

En su trabajo, Shannon buscaba una medida de información en base a la distribución de probabilidad de una variable aleatoria, esto es, el conjunto de valores posibles que puede tomar la variable junto con su probabilidad asociada. Shannon plantea una función H de una distribución de probabilidad $\mathbf{H}(\mathbf{p}_1, \dots, \mathbf{p}_N) = \mathbf{h} \in \mathfrak{R}$ y que debe cumplir las siguientes condiciones:

- 1) H es continua sobre los \mathbf{p}_i ,
- 2) Si los \mathbf{p}_i son iguales, esto es $\mathbf{p}_i = \frac{1}{N}, \forall i$, entonces H crece monótonamente cuando N crece.
- 3) Si una opción puede descomponerse en subdecisiones, entonces H debe ser la suma ponderada de las subopciones.

El Teorema de Shannon establece que la única función que cumple con estas tres condiciones tiene la forma

$$\mathbf{H} = -\mathbf{K} \sum_{i=1}^N \mathbf{p}_i \log_b(\mathbf{p}_i)$$

escogiendo $\mathbf{K}=1$ y $b=2$ se tiene la entropía

$$\mathbf{H} = -\sum_{i=1}^N \mathbf{p}_i \log_2(\mathbf{p}_i)$$

y cuya medida es el bit (de **binary digit**).

Las siguientes propiedades evidencian el significado de la entropía.

Propiedad 1: H es igual a cero si todos los \mathbf{p}_i son igual a cero, excepto uno que tendrá un valor igual a uno.

Propiedad 2: H es máximo e igual a $\log_2(N)$ si los \mathbf{p}_i son iguales, esto es $\mathbf{p}_i = \frac{1}{N}, \forall i$.

Así, la entropía es igual a cero si un evento dentro de la gama posible de eventos ocurre con probabilidad igual a 1, y la entropía es máxima si todos los eventos posibles tienen la misma probabilidad de ocurrir.

Entonces, es evidente que la **entropía es una medida de la incertidumbre** de un sistema representado por el resultado de una variable aleatoria.

Otra medida principal en teoría de la información es aquella de **información mutua**, la cual es una medida de dependencia entre variables aleatorias.

La información mutua [58] de dos variables aleatorias X e Y corresponde a

$$\mathbf{I}(\mathbf{x}; \mathbf{y}) = \mathbf{I}(\mathbf{y}; \mathbf{x}) = \mathbf{H}(\mathbf{x}) + \mathbf{H}(\mathbf{y}) - \mathbf{H}(\mathbf{x}, \mathbf{y})$$

donde $\mathbf{H}(\mathbf{x}, \mathbf{y}) = -\sum \mathbf{p}(\mathbf{x}, \mathbf{y}) \log_2(\mathbf{p}(\mathbf{x}, \mathbf{y}))$ es la entropía del par (X, Y) .

Así, la información mutua es la diferencia entre aquella incertidumbre que cada variable proporciona por separado y aquella que proporcionan de manera conjunta.

En este trabajo se utilizaron dos medidas de asociación entre variables basadas en teoría de la información para medir la relevancia que tiene cada variable sobre la clase respuesta:

- La primera corresponde al coeficiente de incertidumbre simétrico [59] (o coefficient of uncertainty), la cual viene dada por:

$$U(\mathbf{X}, \mathbf{Y}) = 2 \left(\frac{I(\mathbf{X}; \mathbf{Y})}{H(\mathbf{X}) + H(\mathbf{Y})} \right)$$

Esta medida es igual a cero si ambas variables son independientes, esto es, $H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y})$ y es igual a uno si ambas variables son completamente dependientes, esto es, $H(\mathbf{X}) = H(\mathbf{Y}) = H(\mathbf{X}, \mathbf{Y})$.

- La segunda medida utilizada corresponde a la razón entre la entropía total de cada variable X y la suma de la entropía de la misma variable dentro de cada grupo (esto es, la incertidumbre dentro de los modelos correctos y de los modelos incorrectos). Sea X_{-1} la variable X en el espacio de los modelos incorrectos, y X_1 la variable X en el espacio de los modelos correctos. Así, la medida es:

$$M(\mathbf{X}) = \frac{H(\mathbf{X})}{1 + H(\mathbf{X}_{-1}) + H(\mathbf{X}_1)}$$

El uso de medidas de asociación basadas en teoría de la información permite detectar dependencias no lineales entre las variables. En el caso de las medidas clásicas de correlación, sólo detectan dependencias lineales [60].

2.5) Métodos de selección de variables.

Si bien los métodos de ranking de variables proveen información útil en cuanto a la relevancia de cada variable con la clase respuesta, éstos no consideran la redundancia que cada variable tiene con las otras.

Los métodos de selección de variables buscan considerar tanto redundancia como relevancia, y constan de cuatro etapas principales [61], las cuales se describen a continuación.

- 1) Determinación de un punto de partida.

Dado que es posible establecer un orden en el espacio de búsqueda sobre las variables, esto se puede utilizar para definir un punto de partida en el proceso de selección (Fig 13). Cuando la selección se inicia considerando el conjunto completo de variables, se conoce como selección por eliminación reversa, o backward elimination. Por el contrario, cuando la selección se inicia con un conjunto vacío, el proceso se conoce como selección hacia adelante, o forward selection.

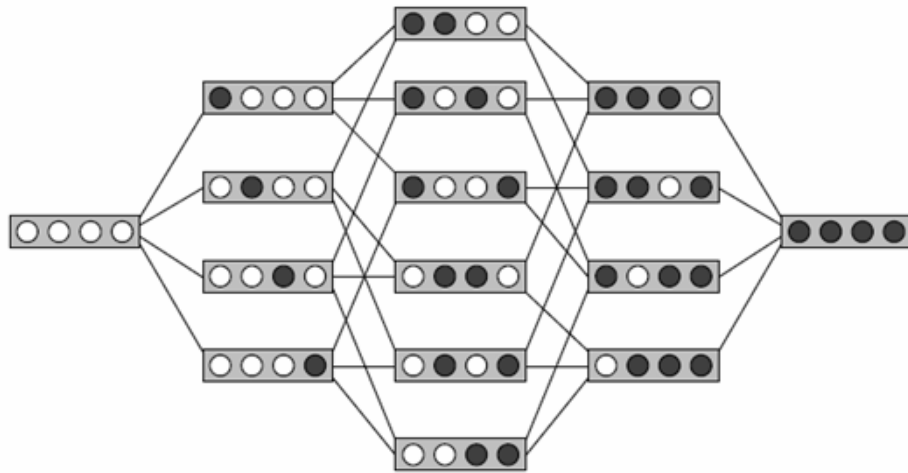


Fig 13. Orden en el espacio de búsqueda de variables. Los puntos negros representan variables seleccionadas; los puntos blancos representan variables descartadas. El inicio de la selección de variables puede darse a partir de un conjunto vacío (de izquierda a derecha), lo que se conoce como forward selection, o desde el conjunto de todas las variables (de derecha a izquierda), lo que se conoce como backward elimination.

2) Organización de la búsqueda

Se refiere a la forma en que se recorre el espacio para determinar el subconjunto óptimo de variables. Para casos en que el número de variables es grande, la búsqueda exhaustiva se vuelve impracticable, pues el tiempo crece exponencialmente con el número de variables, y se debe recurrir a otros métodos tales como búsquedas *greedy* (buscando el óptimo local en cada iteración) u otras alternativas como *best-first search* [62]. Es en esta etapa donde se decide el número de variables a agregar/descartar en cada iteración, de acuerdo al método utilizado.

3) Estrategia a utilizar para evaluar distintos subconjuntos candidatos

En esta etapa se define la estrategia para decidir qué subconjunto de las variables es aquél óptimo. Un criterio común es elegir de acuerdo a la máxima relevancia de las variables con la clase respuesta y la mínima redundancia entre las variables, aunque otras estrategias incluso involucran algoritmos de aprendizaje de manera de maximizar la precisión en cuanto a la clasificación de los individuos (en este trabajo, los modelos proteicos) en sus distintas clases.

4) Uso de un criterio de detención.

Es la última etapa y corresponde a la detención de la búsqueda de acuerdo a algún criterio que puede ser un número fijo de variables (menor al número de variables originales), un número de iteraciones en los cuales la estrategia no sobrepase algún umbral, u otro criterio.

Si bien existen muchos métodos de selección de variables que se pueden utilizar y que han sido descritos en la literatura [63-82], en este trabajo se decidió utilizar dos que se consideran confiables y razonables en lo que se refiere a obtener subconjuntos de variables óptimos en términos de máxima relevancia con respecto a la clase respuesta y

de mínima redundancia entre las variables. Los algoritmos seleccionados se describen a continuación.

2.5.1) Algoritmo de Koller y Sahami

El algoritmo de Koller y Sahami [83] se basa en medidas de teoría de la información para efectuar la selección de variables.

Sea $\mathbf{f} = (\mathbf{f}_1, \dots, \mathbf{f}_n)$ una instancia de un conjunto de variables $\mathbf{F} = (\mathbf{F}_1, \dots, \mathbf{F}_n)$. En este caso, esto corresponde a una fila de la base de datos (sin considerar la clase respuesta) vista como un vector de tamaño 31, con un valor por cada variable medida.

Para cada \mathbf{f} podemos definir $\mathbf{P}(\mathbf{C} | \mathbf{F} = \mathbf{f})$ como la distribución de la clase respuesta \mathbf{C} dado \mathbf{f} .

Sea $\mathbf{G} \subset \mathbf{F}$ y sea \mathbf{f}_g una instancia de \mathbf{G} . La distribución de la clase respuesta, \mathbf{C} dado \mathbf{f}_g está definido por $\mathbf{P}(\mathbf{C} | \mathbf{G} = \mathbf{f}_g)$.

El objetivo es seleccionar \mathbf{G} tal que la distancia entre $\mathbf{P}(\mathbf{C} | \mathbf{F} = \mathbf{f})$ y $\mathbf{P}(\mathbf{C} | \mathbf{G} = \mathbf{f}_g)$ sea mínima.

Sean μ y σ dos distribuciones sobre un espacio de probabilidad Ω . La **entropía cruzada**, o distancia de Kullback y Leibler [84], es una medida de la distancia entre estas dos distribuciones y se define como

$$\mathbf{D}(\mu, \sigma) = \sum_{\mathbf{x} \in \Omega} \mu(\mathbf{x}) \log \left(\frac{\mu(\mathbf{x})}{\sigma(\mathbf{x})} \right).$$

El rol de μ y σ no es simétrico. Si se considera que μ es la distribución “correcta” y σ una aproximación a μ , entonces $\mathbf{D}(\mu, \sigma)$ mide el error de aproximar μ con σ . Sea $\mathbf{P}(\mathbf{C} | \mathbf{F} = \mathbf{f})$ la distribución correcta y $\mathbf{P}(\mathbf{C} | \mathbf{G} = \mathbf{f}_g)$ la aproximación, y dado que el espacio de probabilidad corresponde al conjunto de todas las clasificaciones posibles $\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ (que en este caso es binario) entonces se define

$$\delta_g(\mathbf{f}) = \mathbf{D}(\mathbf{P}(\mathbf{C} | \mathbf{F} = \mathbf{f}), \mathbf{P}(\mathbf{C} | \mathbf{G} = \mathbf{f}_g))$$

como una medida de distancia entre ambas distribuciones. Dado que cada instancia \mathbf{f} puede ocurrir con distinta probabilidad, $\mathbf{P}(\mathbf{f})$, una corrección a la distancia propuesta y que considera esta probabilidad se define como

$$\Delta_g = \sum_{\mathbf{f}} \mathbf{P}(\mathbf{f}) \delta_g(\mathbf{f})$$

Para plantear el algoritmo, se requieren las siguientes definiciones y teoremas.

Def. 1. Independencia Condicional

Dos variables son condicionalmente independientes dado un conjunto de valores \mathbf{X} si, para cualquier asignación de valores \mathbf{a} , \mathbf{b} y \mathbf{x} de las variables \mathbf{A} , \mathbf{B} y \mathbf{X} respectivamente,

$$P(\mathbf{A} = \mathbf{a} \mid \mathbf{X} = \mathbf{x}, \mathbf{B} = \mathbf{b}) = P(\mathbf{A} = \mathbf{a} \mid \mathbf{X} = \mathbf{x})$$

Esto es, \mathbf{B} no aporta información acerca de \mathbf{A} , más allá de la que ya proporciona \mathbf{X} .

Proposición 1. Sea \mathbf{G} un subconjunto de variables y \mathbf{F}_i una variable en \mathbf{G} . Entonces \mathbf{F}_i es condicionalmente independiente de \mathbf{C} dado $\mathbf{G}' = \mathbf{G} - \{\mathbf{F}_i\}$ si y solo si $\Delta_{\mathbf{G}'} = \Delta_{\mathbf{G}}$.

Esto es, se puede eliminar una variable condicionalmente independiente \mathbf{F}_i de \mathbf{G} sin alejarse de la distribución correcta.

Def. 2. Manto de Markov

Sea \mathbf{M} un conjunto de variables que no contiene a \mathbf{F}_i . \mathbf{M} es un manto de Markov para \mathbf{F}_i si \mathbf{F}_i es condicionalmente independiente de $\mathbf{F} - \mathbf{M} - \mathbf{F}_i$, dado \mathbf{M} .

Corolario. Sea \mathbf{G} un subconjunto de variables y \mathbf{F}_i una variable en \mathbf{G} . Sea \mathbf{M} un subconjunto de \mathbf{G} y un manto de Markov para \mathbf{F}_i . Entonces $\Delta_{\mathbf{G}'} = \Delta_{\mathbf{G}}$.

Teorema. Sea \mathbf{G} el conjunto actual de variables, y asumamos que alguna variable (previamente eliminada) $\mathbf{F}_i \notin \mathbf{G}$ tiene un manto de Markov en \mathbf{G} . Sea $\mathbf{F}_j \in \mathbf{G}$ una variable que será eliminada basado en un manto de Markov en \mathbf{G} . Entonces \mathbf{F}_i también tiene un manto de Markov en $\mathbf{G} - \{\mathbf{F}_j\}$.

Así, el criterio del manto de Markov elimina variables que son absolutamente innecesarias.

Hasta aquí, se ha visto cómo se puede eliminar una variable \mathbf{F}_i de un conjunto \mathbf{G} encontrando un manto de Markov \mathbf{M} para \mathbf{F}_i . Sin embargo, es posible que no exista un manto de Markov completo para una variable, o que cubra sólo parcialmente la información que contiene esa variable. Para enfrentar este problema, se propone una heurística que funciona de la siguiente manera: iterativamente buscar un conjunto \mathbf{M}_i candidato para cada \mathbf{F}_i , y usar alguna medida para estimar cuán cercano está \mathbf{M}_i de ser un manto de Markov para \mathbf{F}_i . La variable \mathbf{F}_i para la cual \mathbf{M}_i es más cercana a ser un manto de Markov, es eliminada, y el algoritmo se repite. Para efectos del algoritmo, se elegirá como una aproximación del manto de Markov a algún conjunto de K variables que estén fuertemente correlacionadas con \mathbf{F}_i .

¿cómo se mide cuán cercano está un conjunto \mathbf{M}_i de ser un manto de Markov para \mathbf{F}_i ? Si \mathbf{M}_i es realmente un manto de Markov para \mathbf{F}_i , entonces

$$D(\mathbf{P}(\mathbf{C} | \mathbf{M} = \mathbf{f}_M, \mathbf{F}_i = \mathbf{f}_i), \mathbf{P}(\mathbf{C} | \mathbf{M} = \mathbf{f}_M)) = 0$$

para cualquier asignación de valores \mathbf{f}_M y \mathbf{f}_i a \mathbf{M} y \mathbf{F}_i . Así, se define la entropía cruzada esperada como

$$\delta_G(\mathbf{F}_i | \mathbf{M}_i) = \sum_{\mathbf{f}_{M_i}, \mathbf{f}_i} \mathbf{P}(\mathbf{M}_i = \mathbf{f}_{M_i}, \mathbf{F}_i = \mathbf{f}_i) \cdot D(\mathbf{P}(\mathbf{C} | \mathbf{M} = \mathbf{f}_M, \mathbf{F}_i = \mathbf{f}_i), \mathbf{P}(\mathbf{C} | \mathbf{M} = \mathbf{f}_M))$$

Y por lo tanto $\delta_G(\mathbf{F}_i | \mathbf{M}_i) = 0$ si \mathbf{M}_i es un manto de Markov para \mathbf{F}_i .

Luego, el algoritmo propuesto por Koller y Sahami es el siguiente:

- 1) Se inicia calculando el factor de correlación

$$\rho_{ij} = \frac{\text{Cov}(\mathbf{F}_i, \mathbf{F}_j)}{\text{StDev}(\mathbf{F}_i) \cdot \text{StDev}(\mathbf{F}_j)}$$

para cada par de variables \mathbf{F}_i y \mathbf{F}_j . Luego se instancia $\mathbf{G} = \mathbf{F}$ y se iteran los siguientes pasos hasta que un número predeterminado de variables haya sido eliminado:

- a) Para cada $\mathbf{F}_i \in \mathbf{G}$, tomar \mathbf{M}_i como el conjunto de K variables \mathbf{F}_j en $\mathbf{G} - \{\mathbf{F}_i\}$ para los cuales ρ_{ij} es máximo.
- b) Calcular $\delta_G(\mathbf{F}_i | \mathbf{M}_i)$ para todo i .
- c) Elegir i tal que $\delta_G(\mathbf{F}_i | \mathbf{M}_i)$ es mínimo, y definir $\mathbf{G} = \mathbf{G} - \{\mathbf{F}_i\}$.

2.5.2) Algoritmo de Yu y Liu

Sea un conjunto de datos S , con N variables y una clase respuesta C . Sea

$$SU(\mathbf{X}, \mathbf{Y}) = 2 \left(\frac{I(\mathbf{X}; \mathbf{Y})}{H(\mathbf{X}) + H(\mathbf{Y})} \right)$$

el coeficiente de incertidumbre simétrico. Sea $SU_{i,c}$ el coeficiente de incertidumbre simétrico que mide la correlación entre una variable \mathbf{F}_i y la clase C , y $SU_{j,i}$ la correspondiente medida de correlación entre dos variables \mathbf{F}_i y \mathbf{F}_j , con $j \neq i$.

Sea S' un subconjunto de variables relevantes que es definido de acuerdo a un umbral δ tal que $\forall \mathbf{F}_i \in S', 1 \leq i \leq N, SU_{i,c} \geq \delta$.

El algoritmo de Yu y Liu [85] se basa en las siguientes definiciones:

Def. 1. Correlación Predominante.

La correlación entre una variable \mathbf{F}_i ($\mathbf{F}_i \in \mathbf{S}$) y la clase \mathbf{C} es predominante si y sólo si $\mathbf{SU}_{i,c} \geq \delta$, y $\forall \mathbf{F}_j \in \mathbf{S}'$, ($j \neq i$), no existe \mathbf{F}_j tal que $\mathbf{SU}_{j,i} \geq \mathbf{SU}_{i,c}$.

Si existe un \mathbf{F}_j que cumpla la condición para un \mathbf{F}_i , entonces se le llamará un par redundante a \mathbf{F}_i , y se denotará como \mathbf{S}_{P_i} al conjunto de pares redundantes a \mathbf{F}_i . Dado $\mathbf{F}_i \in \mathbf{S}'$ y \mathbf{S}_{P_i} ($\mathbf{S}_{P_i} \neq \phi$), se divide \mathbf{S}_{P_i} en

$$\mathbf{S}_{P_i}^+ = \{ \mathbf{F}_j \mid \mathbf{F}_j \in \mathbf{S}_{P_i}, \mathbf{SU}_{j,c} > \mathbf{SU}_{i,c} \} \text{ y } \mathbf{S}_{P_i}^- = \{ \mathbf{F}_j \mid \mathbf{F}_j \in \mathbf{S}_{P_i}, \mathbf{SU}_{j,c} \leq \mathbf{SU}_{i,c} \}$$

Def. 2. Variable Predominante.

Una variable es predominante para la clase, si y sólo si la correlación de la variable con la clase es predominante o se puede volver predominante después de remover sus pares redundantes.

De acuerdo a estas dos definiciones, se definen tres heurísticas que juntas pueden identificar variables predominantes y remover variables redundantes entre las relevantes. La idea es que si dos variables son redundantes entre ellas, y se debe eliminar una, se elimina la menos relevante para la clase respuesta.

Heurística 1. (si $\mathbf{S}_{P_i}^+ = \phi$) Tratar \mathbf{F}_i como una variable predominante, remover todas las variables en $\mathbf{S}_{P_i}^-$, y saltar la identificación de pares redundantes para ellas.

Heurística 2. (si $\mathbf{S}_{P_i}^+ \neq \phi$) Procesar todas las variables en $\mathbf{S}_{P_i}^+$ antes de tomar una decisión sobre \mathbf{F}_i . Si ninguna de esas variables se transforma en predominante, seguir Heurística 1; si no sólo remover \mathbf{F}_i y decidir si remover o no variables en $\mathbf{S}_{P_i}^-$ basado en las variables en \mathbf{S}' .

Heurística 3. (punto de partida) La variable con el mayor $\mathbf{SU}_{i,c}$ es siempre una variable predominante y puede ser un punto de partida para remover otras variables.

Basado en estas tres heurísticas, el algoritmo que se utiliza es el que se muestra a continuación:

Input: $\mathbf{S}(\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_N, \mathbf{C})$ //conjunto de variables con su clase respuesta
 δ //umbral predefinido

Output: \mathbf{S}_{best} //un subconjunto óptimo

```

1      begin
2      .   for  $\mathbf{i} = 1$  to  $N$  do begin
3      .           calculate  $\mathbf{SU}_{i,c}$  for  $\mathbf{F}_i$ ;
4      .           if( $\mathbf{SU}_{i,c} \geq \delta$ )

```

```

5           .           append  $F_i$  to  $S'_{list}$ ;
6           .   end;
7           .   order  $S'_{list}$  in descending  $SU_{i,c}$  value;
8           .    $F_p = \text{getFirstElement}(S'_{list})$ ;
9           .   do begin
10          .            $F_q = \text{getNextElement}(S'_{list}, F_p)$ ;
11          .           if ( $F_q \neq \text{NULL}$ )
12          .               do begin
13          .                    $F'_q = F_q$ ;
14          .                   if ( $SU_{p,q} \geq SU_{q,c}$ )
15          .                       remove  $F_q$  from  $S'_{list}$ ;
16          .                        $F_q = \text{getNextElement}(S'_{list}, F_q)$ ;
17          .                       else  $F_q = \text{getNextElement}(S'_{list}, F_q)$ 
18          .                   end until ( $F_q == \text{NULL}$ )
19          .                    $F_p = \text{getNextElement}(S'_{list}, F_p)$ ;
20          .               end until ( $F_p == \text{NULL}$ )
21          .    $S_{best} = S'_{list}$ ;
22          . end;

```

2.6) Métodos de extracción de variables.

Los métodos de extracción de variables son todos aquellos que combinan las variables originales para dar origen a nuevas variables, de manera que éstas capturen la mayor cantidad de información posible. Así, metodologías tan diversas como redes auto asociativas [86], análisis de componentes independientes [87], algoritmos genéticos [88] (aunque su rango de aplicabilidad es mucho más general) y análisis de componentes principales [37] se pueden utilizar para extraer nuevas variables de manera de reducir la dimensionalidad del problema y extraer tanta información como sea posible de las variables originales.

En este trabajo se utilizó el método de análisis de componentes principales (o ACP), que se explica a continuación.

2.6.1) Análisis de componentes principales

Son muchas las situaciones en las cuales más de 3 variables son medidas, y por lo tanto el uso de herramientas gráficas que permitan visualizar el comportamiento de estas variables no es posible.

Con el fin de reducir la dimensionalidad de los datos y a la vez capturar tanta información como sea posible de las variables originales, el método de análisis de

componentes principales genera nuevas variables – llamadas componentes principales – que son una combinación lineal de las variables originales.

Para ilustrar la técnica, se considera el caso en que se tienen puntos correspondientes a los valores de dos variables, \mathbf{x}_1 y \mathbf{x}_2 . Este conjunto de datos se puede representar en un nuevo conjunto de coordenadas $(\mathbf{c}_1, \mathbf{c}_2)$ llamado componentes principales (Fig 14).

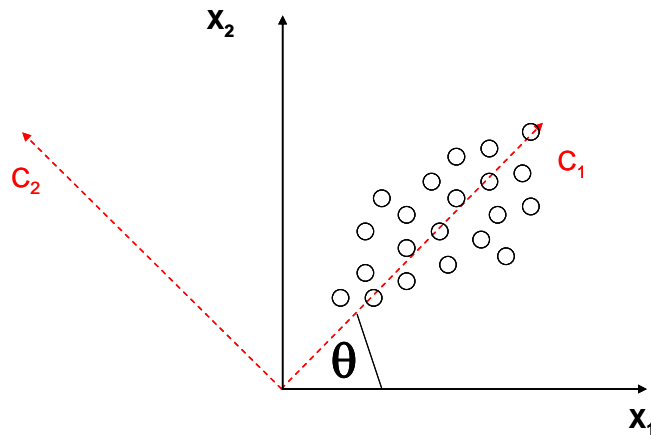


Fig 14. Componentes principales sobre los datos. Las componentes principales corresponden a un nuevo conjunto de coordenadas que maximizan la dispersión de los datos sobre estos, de manera de capturar la mayor cantidad de información posible en el menor número de coordenadas posible.

Este nuevo par de ejes $(\mathbf{c}_1, \mathbf{c}_2)$ permite que la dispersión de los datos proyectados sobre \mathbf{c}_1 es máxima con respecto a cualquier otro eje.

La relación entre las variables $(\mathbf{x}_1, \mathbf{x}_2)$ y $(\mathbf{c}_1, \mathbf{c}_2)$ queda establecida por las siguientes ecuaciones:

$$\begin{aligned} \mathbf{c}_1 &= \cos(\theta) \cdot \mathbf{x}_1 + \text{sen}(\theta) \cdot \mathbf{x}_2 \\ \mathbf{c}_2 &= -\text{sen}(\theta) \cdot \mathbf{x}_1 + \cos(\theta) \cdot \mathbf{x}_2 \end{aligned}$$

Esto se puede escribir matricialmente si se define

$$\mathbf{U} = \begin{bmatrix} \cos(\theta) & \text{sen}(\theta) \\ -\text{sen}(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} \mathbf{u}_{11} & \mathbf{u}_{12} \\ \mathbf{u}_{21} & \mathbf{u}_{22} \end{bmatrix}$$

$$\bar{\mathbf{c}} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix}, \bar{\mathbf{x}} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$$

quedando

$$\bar{\mathbf{C}} = \mathbf{U} \cdot \bar{\mathbf{X}}$$

De esta forma, es evidente que \mathbf{U} es una matriz de rotación aplicada a los ejes $(\mathbf{x}_1, \mathbf{x}_2)$. Entonces, el objetivo del ACP es encontrar aquella matriz \mathbf{U} que maximiza la dispersión de los datos sobre los nuevos ejes.

Para presentar la técnica, se generaliza para p variables $(\mathbf{x}_1, \dots, \mathbf{x}_p)$. En este caso

$$\underset{p \times 1}{\bar{\mathbf{C}}} = \underset{p \times p}{\mathbf{U}} \cdot \underset{p \times 1}{\bar{\mathbf{X}}}$$

Sea $\mathbf{VAR}(\bar{\mathbf{X}}) = \Gamma$ la matriz de varianza y covarianza para $\bar{\mathbf{X}}$.

El problema se plantea como

$$\begin{aligned} & \underset{\mathbf{U}}{\text{MAX}} \{ \mathbf{VAR}(\bar{\mathbf{C}}) \} \\ \text{s.a.} & \mathbf{U}^t \mathbf{U} = \mathbf{I} \end{aligned}$$

donde \mathbf{I} es la matriz de identidad.

La condición $\mathbf{U}^t \mathbf{U} = \mathbf{I}$ se impone porque se busca que los \mathbf{C}_i sean ortogonales entre sí. Esto se cumple si los $\mathbf{u}_{\bullet i}$ (las columnas de la matriz) son ortogonales entre sí, pues estos vectores definen las direcciones de las componentes principales.

Para resolver el problema basta plantear el lagrangeano como sigue:

$$\begin{aligned} \mathbf{L} &= \mathbf{VAR}(\bar{\mathbf{C}}) - \lambda \cdot (\mathbf{U}^t \mathbf{U} - \mathbf{I}) = 0 \\ &\Leftrightarrow \\ \mathbf{L} &= \mathbf{U} \Gamma \mathbf{U}^t - \lambda \cdot (\mathbf{U}^t \mathbf{U} - \mathbf{I}) = 0 \end{aligned}$$

Aplicando cálculo matricial, se tiene que

$$\frac{d\mathbf{L}}{d\mathbf{U}} = 0 \Rightarrow (\Gamma - \lambda \cdot \mathbf{I}) \cdot \mathbf{U} = 0 \Rightarrow |\Gamma - \lambda \cdot \mathbf{I}| = 0$$

Lo que corresponde a la ecuación característica de la matriz Γ . Las soluciones a esta ecuación (cada λ_i) son los valores propios de Γ y para cada λ_i se tiene un vector propio $\mathbf{u}_{\bullet i}$ asociado. Además, se cumplen las siguientes propiedades:

- Dado que Γ es una matriz simétrica, entonces todos sus valores propios son reales.
- Dado que Γ es definida positiva, entonces sus valores propios están ordenados.

Adicionalmente se puede demostrar que $\text{VAR}(\bar{\mathbf{C}}) = \mathbf{U}\Gamma\mathbf{U}^t$ es una matriz diagonal formada por los valores propios de Γ , por lo tanto, los valores propios corresponden a las varianzas de los nuevos ejes.

Dado que los λ_i están ordenados, y ya que cada λ_i tiene asociado un vector propio $\mathbf{u}_{\cdot i}$ que define la dirección del eje i , entonces se puede establecer un orden entre las nuevas variables, de manera que:

- \mathbf{C}_1 es el primer componente principal, y como se cumple que $\lambda_1 > \lambda_i \quad \forall i \neq 1$, entonces la dirección de la máxima varianza de los datos está determinada por \mathbf{C}_1 .
- \mathbf{C}_2 es la segunda componente principal, entonces la dirección de máxima varianza entre todos los ejes ortogonales a \mathbf{C}_1 es \mathbf{C}_2 , y así sucesivamente para las p componentes principales.

Como la idea del método es proporcionar un conjunto de nuevas variables $\mathbf{m} < \mathbf{p}$, entonces se debe establecer un criterio de corte al número de componentes principales a considerar. Por lo general ese criterio está dado por un número \mathbf{h} de componentes de manera que se cumpla:

$$\frac{\sum_{i=1}^{\mathbf{h}} \lambda_i}{\sum_{i=1}^{\mathbf{p}} \lambda_i} \geq 0.9$$

Esto es, que las primeras \mathbf{h} componentes principales capturen al menos el 90% de la variabilidad presente en los datos.

Cabe señalar que en el caso que se trabaje con los datos centrados y reducidos, la matriz de varianza y covarianza se transforma en la matriz de correlaciones, en la cual se tiene para cada elemento (\mathbf{i}, \mathbf{j}) de la matriz, la correlación entre las variables \mathbf{x}_i y \mathbf{x}_j .

2.7) Algoritmos de aprendizaje y clasificación.

Una vez detectados los modelos outliers y seleccionadas aquellas variables que son más relevantes para la clase respuesta y menos redundantes entre sí, el paso siguiente es utilizar diversos algoritmos de aprendizaje para generar clasificadores, y aplicar estos clasificadores para medir el rendimiento sobre un conjunto de datos destinado para ese fin. A continuación se formaliza la diferencia entre un algoritmo de aprendizaje y un clasificador [89].

Sea un conjunto \mathbf{X} de puntos posibles de datos, llamado **población**. Sea una **función objetivo**, \mathbf{f} que clasifica cada $\mathbf{x} \in \mathbf{X}$ en una de k clases. Para efectos de este trabajo y sin perder generalidad se define $k=2$ para los modelos correctos e incorrectos.

Sea una **muestra** \mathcal{S} tomada al azar de \mathbf{x} . Un **conjunto de entrenamiento** se construye etiquetando cada $\mathbf{x} \in \mathcal{S}$ de acuerdo a $\mathbf{f}(\mathbf{x})$. Así, cada modelo en el conjunto de entrenamiento se puede entender como un par $\langle \mathbf{x}, \mathbf{f}(\mathbf{x}) \rangle$.

Un **algoritmo de aprendizaje** \mathbf{A} toma como entrada un conjunto de entrenamiento \mathbf{R} y arroja como salida un clasificador $\hat{\mathbf{f}}$. Un **clasificador** $\hat{\mathbf{f}}$ toma como entrada un conjunto de modelos sin etiquetar, y arroja como salida la etiquetación para cada uno.

La **tasa de error verdadera** del clasificador $\hat{\mathbf{f}}$ corresponde a la probabilidad de que este clasifique mal un modelo tomado al azar del conjunto \mathbf{x} . En la práctica, esta tasa de error no se conoce y lo que se hace es calcular la **tasa de error estimada**, lo que se logra tomando una muestra \mathcal{S} y subdividiendo esta muestra en un **conjunto de entrenamiento** \mathbf{R} y un **conjunto de prueba** \mathbf{T} . La tasa de error de $\hat{\mathbf{f}}$ sobre \mathbf{T} da una estimación de la tasa de error verdadera de $\hat{\mathbf{f}}$ sobre la población \mathbf{x} .

2.7.1) Variables simples como clasificadores

El clasificador más sencillo se puede obtener de cada una de las variables medidas sobre los modelos. Dado que cada variable toma un rango continuo de valores, y dado que cada modelo correcto e incorrecto toma un valor para esa variable (a ese valor le llamaremos *score*) entonces los modelos correctos se distribuyen de una forma y los incorrectos de otra, al menos que las distribuciones sean idénticas (Fig 15).

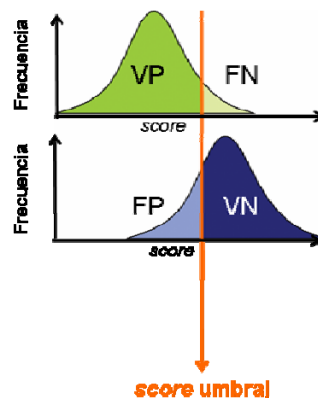


Fig 15. Esquema de la distribución de modelos correctos e incorrectos para una variable. La distribución de los modelos correctos (en verde) e incorrectos (en azul) para una variable cualquiera que actúa como *score* permite determinar un umbral (flecha vertical) que separa ambas clases de manera de maximizar el número de modelos correctamente clasificados en cada una de sus categorías. FN: Falsos Negativos, FP: Falsos Positivos, VP: Verdaderos Positivos, VN: Verdaderos Negativos.

Así, basta recorrer el espacio de valores que toma la variable para encontrar aquél **umbral óptimo** de manera que el número de Falsos Negativos (FN) - esto es, modelos correctos clasificados como incorrectos - y de Falsos Positivos (FP) - esto es, modelos incorrectos clasificados como correctos - sea mínimo.

Para determinar el umbral óptimo, se utiliza el conjunto de entrenamiento \mathbf{R} , y luego se mide cuantos FN y FP se contabilizan al aplicar ese umbral óptimo sobre el conjunto de prueba, \mathbf{T} .

2.7.2) Mínima distancia a centroides.

Las n variables medidas para cada uno de los modelos correctos e incorrectos define un vector $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ que se puede situar en \mathcal{R}^n (Fig 16).

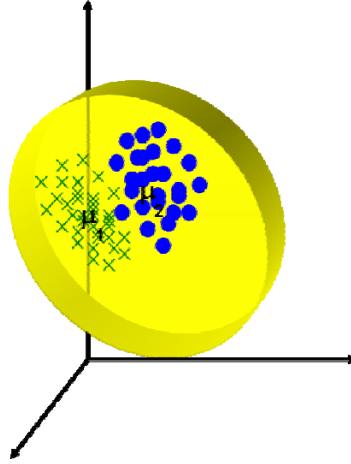


Fig 16. Esquema de la distribución de modelos correctos e incorrectos en el espacio n-dimensional. En este esquema hipotético, cada cruz representa un modelo correcto, y cada punto un modelo incorrecto. μ_1 y μ_2 corresponden a los centroides de los modelos correctos e incorrectos, respectivamente.

Sea $\bar{\mu}_1 = (\mu_{11}, \mu_{12}, \dots, \mu_{1n})$ el centroide de los modelos correctos y $\bar{\mu}_2 = (\mu_{21}, \mu_{22}, \dots, \mu_{2n})$ el centroide de los modelos incorrectos, entonces se puede utilizar el conjunto de entrenamiento R para estimar estos dos parámetros de la siguiente manera:

$$\bar{\mu}_1 \approx \hat{\mu}_1 = \left(\frac{1}{m} \sum_{i=1}^m \mathbf{x}_{11}^i, \frac{1}{m} \sum_{i=1}^m \mathbf{x}_{12}^i, \dots, \frac{1}{m} \sum_{i=1}^m \mathbf{x}_{1n}^i \right)$$

$$\bar{\mu}_2 \approx \hat{\mu}_2 = \left(\frac{1}{s} \sum_{i=1}^s \mathbf{x}_{21}^i, \frac{1}{s} \sum_{i=1}^s \mathbf{x}_{22}^i, \dots, \frac{1}{s} \sum_{i=1}^s \mathbf{x}_{2n}^i \right)$$

Donde m y s corresponden al número de modelos correctos e incorrectos en el conjunto de entrenamiento, respectivamente.

Luego, se clasifica un modelo $\bar{\mathbf{x}} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ del conjunto de prueba T como correcto si

$$\|\bar{\mathbf{x}} - \hat{\mu}_1\| \leq \|\bar{\mathbf{x}} - \hat{\mu}_2\|$$

o como incorrecto si

$$\|\bar{\mathbf{x}} - \hat{\mu}_1\| > \|\bar{\mathbf{x}} - \hat{\mu}_2\|$$

2.7.3) Naive Bayes

Naive Bayes es uno de los clasificadores más simples y una alternativa bayesiana al problema de clasificación. Sean 2 eventos A y B, entonces la probabilidad de A dado B está dada por:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

De la misma forma,

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

Reemplazando ambas ecuaciones se obtiene que

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

De manera equivalente, se puede establecer la probabilidad de la clase respuesta C dado un vector de variables $\vec{x} = (x_1, \dots, x_n)$ como

$$P(C | x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n | C) \cdot P(C)}{P(x_1, \dots, x_n)}$$

Lo “naive” (ingenuo) de este método radica en suponer que las variables (x_1, \dots, x_n) son estadísticamente independientes, esto es

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i)$$

Así, la probabilidad de cada clase dada la observación $\vec{x} = (x_1, \dots, x_n)$ está dada por

$$P(C | x_1, \dots, x_n) = \frac{\prod_{i=1}^n P(x_i | C) \cdot P(C)}{\prod_{i=1}^n P(x_i)}$$

esto es, dada una observación, se puede calcular la probabilidad de que la observación pertenezca a cada una de las clases posibles, que en este caso corresponde a modelos correctos o incorrectos. Así, sea C_1 la clase de los modelos correctos y C_2 la clase de modelos incorrectos, un modelo $\vec{x} = (x_1, \dots, x_n)$ es clasificado como correcto si

$$P(C_1 | x_1, \dots, x_n) > P(C_2 | x_1, \dots, x_n) \Leftrightarrow \prod_{i=1}^n P(x_i | C_1) \cdot P(C_1) > \prod_{i=1}^n P(x_i | C_2) \cdot P(C_2)$$

o como incorrecto si

$$\mathbf{P}(\mathbf{C}_1 | \mathbf{x}_1, \dots, \mathbf{x}_n) \leq \mathbf{P}(\mathbf{C}_2 | \mathbf{x}_1, \dots, \mathbf{x}_n) \Leftrightarrow \prod_{i=1}^n \mathbf{P}(\mathbf{x}_i | \mathbf{C}_1) \cdot \mathbf{P}(\mathbf{C}_1) \leq \prod_{i=1}^n \mathbf{P}(\mathbf{x}_i | \mathbf{C}_2) \cdot \mathbf{P}(\mathbf{C}_2)$$

La estimación de $\mathbf{P}(\mathbf{C})$ para cada clase se logra calculando la frecuencia del número de casos correctos e incorrectos en el conjunto de entrenamiento y dividiendo por el total de casos.

La estimación de $\mathbf{P}(\mathbf{x}_i | \mathbf{C})$ para cada variable \mathbf{x}_i se puede obtener de dos formas para el caso de variables continuas:

- discretizando el rango de valores que toma cada variable \mathbf{x}_i . Así, de acuerdo al intervalo de valores en que cae \mathbf{x}_i , se estima su probabilidad de acuerdo a la frecuencia de ocurrencia (Fig 17).

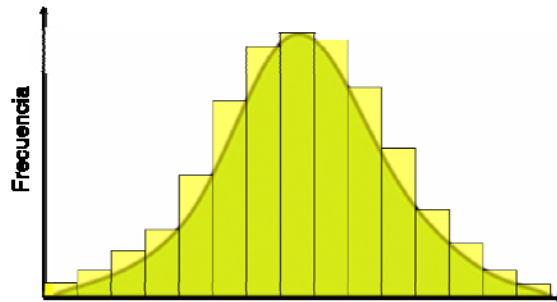


Fig 17. Esquema de discretización del rango de valores de una variable continua. Para el rango completo de valores que toma la variable continua, se define un número finito de intervalos (barras en la figura).

- asumiendo que cada variable se distribuye como una normal. Si llamamos $\bar{\boldsymbol{\mu}}_1 = (\boldsymbol{\mu}_{11}, \boldsymbol{\mu}_{12}, \dots, \boldsymbol{\mu}_{1n})$ y $\bar{\boldsymbol{\sigma}}_1^2 = (\boldsymbol{\sigma}_{11}^2, \dots, \boldsymbol{\sigma}_{1n}^2)$ a la esperanza y varianza de los modelos correctos, $\bar{\boldsymbol{\mu}}_2 = (\boldsymbol{\mu}_{21}, \boldsymbol{\mu}_{22}, \dots, \boldsymbol{\mu}_{2n})$ y $\bar{\boldsymbol{\sigma}}_2^2 = (\boldsymbol{\sigma}_{21}^2, \dots, \boldsymbol{\sigma}_{2n}^2)$ a la esperanza y varianza de los modelos incorrectos, entonces

$$\mathbf{P}(\mathbf{x}_i | \mathbf{C}_j) = \left(\frac{1}{\sqrt{2\pi\boldsymbol{\sigma}_{ji}^2}} \right) \exp\left(-\frac{1}{2\boldsymbol{\sigma}_{ji}^2} (\mathbf{x}_i - \boldsymbol{\mu}_{ji})^2 \right)$$

Utilizando el conjunto de entrenamiento se puede estimar la esperanza y varianza para cada caso como sigue:

$$\bar{\boldsymbol{\mu}}_1 \approx \hat{\boldsymbol{\mu}}_1 = \left(\frac{1}{\mathbf{m}} \sum_{i=1}^{\mathbf{m}} \mathbf{x}_{11}^i, \frac{1}{\mathbf{m}} \sum_{i=1}^{\mathbf{m}} \mathbf{x}_{12}^i, \dots, \frac{1}{\mathbf{m}} \sum_{i=1}^{\mathbf{m}} \mathbf{x}_{1n}^i \right)$$

$$\bar{\boldsymbol{\mu}}_2 \approx \hat{\boldsymbol{\mu}}_2 = \left(\frac{1}{\mathbf{s}} \sum_{i=1}^{\mathbf{s}} \mathbf{x}_{21}^i, \frac{1}{\mathbf{s}} \sum_{i=1}^{\mathbf{s}} \mathbf{x}_{22}^i, \dots, \frac{1}{\mathbf{s}} \sum_{i=1}^{\mathbf{s}} \mathbf{x}_{2n}^i \right)$$

y

$$\begin{aligned}\bar{\sigma}_1^2 &\approx \hat{\sigma}_1^2 = \left(\frac{1}{\mathbf{m}} \sum_{i=1}^{\mathbf{m}} (\mathbf{x}_{11}^i - \bar{\mathbf{x}}_{11})^2, \dots, \frac{1}{\mathbf{m}} \sum_{i=1}^{\mathbf{m}} (\mathbf{x}_{1n}^i - \bar{\mathbf{x}}_{1n})^2 \right) \\ \bar{\sigma}_2^2 &\approx \hat{\sigma}_2^2 = \left(\frac{1}{\mathbf{m}} \sum_{i=1}^{\mathbf{m}} (\mathbf{x}_{21}^i - \bar{\mathbf{x}}_{21})^2, \dots, \frac{1}{\mathbf{m}} \sum_{i=1}^{\mathbf{m}} (\mathbf{x}_{2n}^i - \bar{\mathbf{x}}_{2n})^2 \right)\end{aligned}$$

Donde m y s corresponden al número de modelos correctos e incorrectos en el conjunto de entrenamiento, respectivamente.

Luego, estas estimaciones se reemplazan en el criterio de clasificación, y se evalúa el clasificador utilizando el conjunto de prueba.

2.7.4) Máquinas de vectores de soporte

Las máquinas de vectores de soporte [90-92], o support vector machines (SVM), son un algoritmo de aprendizaje que busca el mejor hiperplano de separación en términos de generalización. Esto es, dado numerosos hiperplanos que son soluciones para separar correctamente ambas clases - suponiendo que son linealmente separables (Fig 18) - elegir aquél que sea más robusto en términos de que represente una solución genérica frente a posibles nuevos puntos en el espacio que sea necesario clasificar. El hiperplano separador óptimo en términos de generalización de acuerdo a SVM es aquél equidistante a los puntos más cercanos a cada clase.

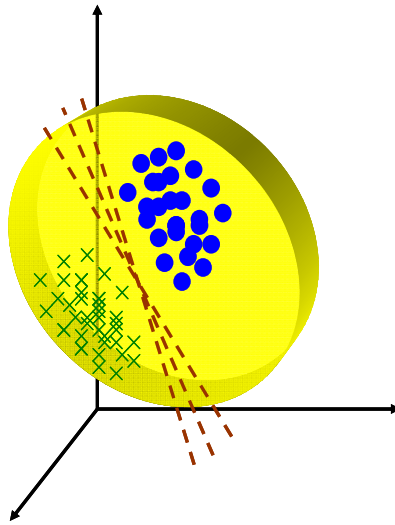


Fig 18. Múltiples hiperplanos separadores. Suponiendo que los modelos correctos, representados por una cruz, y los modelos incorrectos, representados por un punto, son dos clases linealmente separables, entonces existen infinitos hiperplanos separadores que son una solución al problema de la clasificación.

Para formalizar este criterio, a continuación se describe matemáticamente el algoritmo de aprendizaje [93].

Todo hiperplano en un espacio d -dimensional \mathcal{R}^d se puede expresar como $\mathbf{h}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + \mathbf{b}$, donde $\mathbf{w} \in \mathcal{R}^d$ es un vector ortogonal al hiperplano, \mathbf{x} es un vector

que contiene las d variables medidas para cada modelo, $\mathbf{b} \in \mathfrak{R}$ se le llama sesgo, y $\langle \cdot, \cdot \rangle$ corresponde al producto punto entre dos vectores.

Sea γ el **margen geométrico**, correspondiente a la distancia mínima entre los modelos considerados como puntos en el espacio d -dimensional y el hiperplano (Fig19).

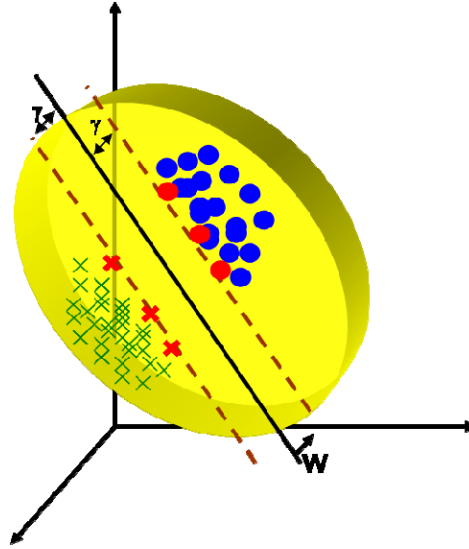


Fig 19. Hiperplano equidistante y margen geométrico. El margen geométrico γ define la distancia de los puntos más cercanos de cada clase al hiperplano equidistante. Los puntos rojos corresponden a los vectores de soporte.

Sea un problema de clasificación binaria dado por un conjunto de n datos $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$ donde cada $\mathbf{x}_i \in \mathfrak{R}^d$ y cada $\mathbf{y}_i \in \{-1, +1\}$ denota si el modelo i es correcto (+1) o incorrecto (-1). A continuación se presentan tres formas de SVM de manera de incluir soluciones no lineales para el problema de clasificación y flexibilidad en el error de clasificación.

2.7.4.1) SVM lineal con margen máximo.

La distancia de un vector \mathbf{x} a un hiperplano \mathbf{h} viene dado por

$$\text{dist}(\mathbf{h}, \mathbf{x}) = \frac{|\mathbf{h}(\mathbf{x})|}{\|\mathbf{w}\|}$$

Donde $\|\mathbf{w}\|$ es la norma de $\mathbf{w} \in \mathfrak{R}^d$. Dado que el conjunto de datos es en teoría linealmente separable, se puede reescalar \mathbf{w} y \mathbf{b} de manera que la distancia de los vectores \mathbf{x} más cercanos al hiperplano sea $\frac{1}{\|\mathbf{w}\|}$. Así, los $\mathbf{x}_i \in \mathfrak{R}^d$ tendrán $|\mathbf{h}(\mathbf{x}_i)| \geq 1$

Con esto, el problema de encontrar el hiperplano equidistante a dos clases se plantea como:

$$\begin{aligned} & \mathbf{MAX} \frac{1}{\|\mathbf{w}\|} \\ & \text{s.a} \\ & \mathbf{y}_i \cdot (\langle \mathbf{w}, \mathbf{x}_i \rangle + \mathbf{b}) \geq 1 \end{aligned}$$

con $1 \leq \mathbf{i} \leq \mathbf{n}$. La expresión $\mathbf{y}_i \cdot (\langle \mathbf{w}, \mathbf{x}_i \rangle + \mathbf{b})$ se conoce como el **margen funcional** de $(\mathbf{x}_i, \mathbf{y}_i)$, y $\frac{1}{\|\mathbf{w}\|}$ corresponde al **margen geométrico** del hiperplano.

La formulación típica del problema corresponde a uno de optimización convexa, consistente en minimizar una función cuadrática bajo restricciones en forma de desigualdad lineal (formulación primal):

$$\begin{aligned} & \mathbf{MIN} \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle \\ & \text{s.a} \\ & \mathbf{y}_i \cdot (\langle \mathbf{w}, \mathbf{x}_i \rangle + \mathbf{b}) \geq 1 \end{aligned}$$

Pero la formulación más utilizada para resolver el problema – pues es más fácil de resolver – corresponde a aquella dual equivalente, y que nos da una solución analítica única y exacta. Esta formulación dual se expresa en el siguiente teorema.

Teorema. Sea un conjunto de datos \mathbf{S} linealmente separable en el espacio de variables. Sea α^* una solución del siguiente problema dual:

$$\begin{aligned} & \mathbf{MAX} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \mathbf{y}_i \mathbf{y}_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ & \text{s.a.} \\ & \sum_{i=1}^n \mathbf{y}_i \alpha_i = 0 \\ & \alpha_i \geq 0 \end{aligned}$$

con $1 \leq \mathbf{i} \leq \mathbf{n}$.

Entonces el vector $\mathbf{w}^* = \sum_{i=1}^n \mathbf{y}_i \alpha_i^* \mathbf{x}_i$ es el vector ortogonal al hiperplano con margen geométrico máximo. La SVM lineal con margen máximo es entonces

$$\mathbf{h}(\mathbf{x}) = \langle \mathbf{w}^*, \mathbf{x} \rangle + \mathbf{b}^* = \sum_{i=1}^n \mathbf{y}_i \alpha_i^* \langle \mathbf{x}_i, \mathbf{x} \rangle + \mathbf{b}^*$$

donde

$$\mathbf{b}^* = -\frac{1}{2} \left(\underset{y_j=-1}{\text{MAX}} \left\{ \langle \mathbf{w}^*, \mathbf{x}_j \rangle \right\} + \underset{y_j=+1}{\text{MIN}} \left\{ \langle \mathbf{w}^*, \mathbf{x}_j \rangle \right\} \right)$$

El clasificador asociado es

$$\mathbf{f}(\mathbf{x}) = \text{signo}(\mathbf{h}(\mathbf{x}))$$

Por un lado, la expresión $\mathbf{w}^* = \sum_{i=1}^n \mathbf{y}_i \alpha_i^* \mathbf{x}_i$ indica que el vector ortogonal al hiperplano es una combinación lineal de los n modelos del conjunto de datos. Por otro lado, la condición de Karush-Kuhn-Tucker necesaria para transformar el problema primal en aquel dual equivalente establece que

$$\alpha_i^* \left(1 - \mathbf{y}_i \left(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + \mathbf{b}^* \right) \right) = 0$$

para $1 \leq i \leq n$, lo que implica que solo algunos α_i^* son iguales a 0. Así, solo algunos modelos formarán parte de la ecuación que conforma \mathbf{w}^* . Aquellos modelos que tienen $\alpha_i^* \neq 0$ son los **vectores de soporte**. Estos modelos se encuentran en la frontera de la región de decisión y tienen margen funcional igual a 1. Ejemplos de vectores de soporte se muestran con color rojo en Fig 19.

2.7.4.2) SVM con margen máximo en el espacio de características.

La formulación planteada hasta ahora sólo resuelve el problema para conjuntos de datos linealmente separables.

Dado un conjunto de datos que no es linealmente separable, se recurre a las **funciones núcleo** (o kernel). Una función núcleo es:

$$\mathbf{K} : \mathcal{R}^d \times \mathcal{R}^d \rightarrow \mathcal{R}$$

tal que $\mathbf{K}(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$, donde $\Phi : \mathcal{R}^d \rightarrow \mathfrak{S}$ es una función que realiza una transformación no lineal de un modelo en \mathcal{R}^d a un espacio de características \mathfrak{S} de mayor dimensión y en el cual el problema de clasificación binaria se vuelve linealmente separable. Así, se puede reformular el problema utilizando las funciones núcleo como sigue:

$$\text{MAX} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \mathbf{y}_i \mathbf{y}_j \alpha_i \alpha_j \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$$

s.a.

$$\sum_{i=1}^n \mathbf{y}_i \alpha_i = 0$$

$$\alpha_i \geq 0$$

con $1 \leq i \leq n$.

Entonces el vector $\mathbf{w}^* = \sum_{i=1}^n \mathbf{y}_i \alpha_i^* \Phi(\mathbf{x}_i)$ es el vector ortogonal al hiperplano con margen geométrico máximo **en el espacio de características**. La SVM con margen máximo en el espacio de características es entonces

$$\mathbf{h}(\mathbf{x}) = \langle \mathbf{w}^*, \Phi(\mathbf{x}) \rangle + \mathbf{b}^* = \sum_{i=1}^n \mathbf{y}_i \alpha_i^* \mathbf{K}(\mathbf{x}_i, \mathbf{x}) + \mathbf{b}^*$$

donde

$$\mathbf{b}^* = -\frac{1}{2} \left(\underset{\mathbf{y}_j=-1}{\text{MAX}} \{ \langle \mathbf{w}^*, \Phi(\mathbf{x}_j) \rangle \} + \underset{\mathbf{y}_j=+1}{\text{MIN}} \{ \langle \mathbf{w}^*, \Phi(\mathbf{x}_j) \rangle \} \right)$$

El clasificador asociado es

$$\mathbf{f}(\mathbf{x}) = \text{signo}(\mathbf{h}(\mathbf{x}))$$

Algunos ejemplos de funciones núcleo son:

Lineal	$\mathbf{x}^t \mathbf{y}$	
Polinómica	$(\langle \mathbf{x}, \mathbf{y} \rangle + \mathbf{c})^d$	$\mathbf{c} \in \mathfrak{R}, \mathbf{d} \in \mathfrak{N}$
Gaussiana	$\exp\left(\frac{-\ \mathbf{x} - \mathbf{y}\ ^2}{\gamma}\right)$	$\gamma > 0$
Sigmoidal	$\tanh(\mathbf{s}\langle \mathbf{x}, \mathbf{y} \rangle + \mathbf{r})$	$\mathbf{s}, \mathbf{r} \in \mathfrak{R}$

2.7.4.3) SVM con margen blando.

Con el fin de evitar un sobreajuste a los datos al momento de encontrar una solución, se puede incluir un término que permita dar cierta holgura a los errores en la clasificación, en pos de una solución más genérica (Fig 20). La adición de variables de holgura (ξ) permite que las restricciones en el problema de optimización no se tengan que cumplir de manera estricta.

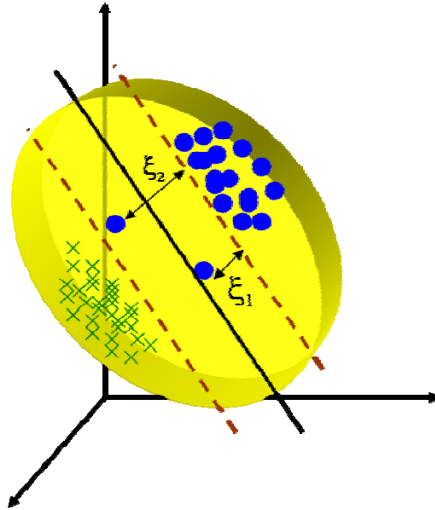


Fig 20. Variables de holgura en el problema de optimización. La inclusión de estas variables de holgura en el problema de optimización planteado por SVM permite encontrar soluciones más generales al permitir cierto error en la clasificación, evitando así un sobreajuste de los datos.

Así, el problema de optimización primal es:

$$\begin{aligned} \text{MIN} \quad & \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle + C \sum_{i=1}^n \xi_i \\ \text{s.a.} \quad & \mathbf{y}_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + \mathbf{b}) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

con $1 \leq i \leq n$, $C > 0$.

C determina la holgura del margen blando y se fija *a priori*. Valores de C grandes obligan a tener pocas variables de holgura diferentes de cero en la solución final.

El problema dual equivalente está dado por:

$$\begin{aligned} \text{MAX} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \mathbf{y}_i \mathbf{y}_j \alpha_i \alpha_j \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.a.} \quad & \sum_{i=1}^n \mathbf{y}_i \alpha_i = 0 \\ & C \geq \alpha_i \geq 0 \end{aligned}$$

con $1 \leq i \leq n$.

La SVM con margen blando en el espacio de características es entonces

$$\mathbf{h}(\mathbf{x}) = \sum_{i=1}^n \mathbf{y}_i \alpha_i^* \mathbf{K}(\mathbf{x}_i, \mathbf{x}) + \mathbf{b}^*$$

Donde \mathbf{b}^* es tal que $\mathbf{y}_i \cdot \mathbf{h}(\mathbf{x}_i) = 1$ para todo i que cumple $0 \leq \alpha_i^* \leq \mathbf{C}$. El clasificador asociado es $\mathbf{f}(\mathbf{x}) = \text{signo}(\mathbf{h}(\mathbf{x}))$.

El espacio de valores posibles y muestreados para los parámetros utilizados en este algoritmo de aprendizaje se muestran en las Tablas 2 y 3.

Tabla 2. Espacio de valores posibles para los parámetros de SVM.

KERNEL	d	γ	r	C
Lineal	N.A.	N.A.	N.A.	\mathcal{R}^+
Polinomial	$\mathcal{N} \cup \{0\}$	\mathcal{R}^+	\mathcal{R}	\mathcal{R}^+
Gaussiana	N.A.	\mathcal{R}^+	N.A.	\mathcal{R}^+
Sigmoidal	N.A.	\mathcal{R}^+	\mathcal{R}	\mathcal{R}^+

El parámetro d representa el exponente en la función kernel polinómica. γ es un factor de ponderación de la medida aplicada sobre dos puntos x e y . r corresponde a un desplazamiento en el espacio. C es el parámetro que controla las variables de holgura.

Tabla 3. Espacio de valores muestreados para los parámetros de SVM.

KERNEL	d	γ	r	C
Lineal	N.A.	N.A.	N.A.	[0.1-100.0] p:0.1
Polinomial	3,4	[0.0-0.9] p:0.1	-10,0,5,10	[0.1-20.0] p:0.1
Gaussiana	N.A.	[0.0-9.9] p:0.1	N.A.	[0.1-10.0] p:0.1
Sigmoidal	N.A.	[0.0-9.9] p:0.1	-20,-15,-10	[0.1-5.0] p:0.1

La notación p: 0.1 indica que la variable en cuestión se varió con un paso de 0,1 desde el menor al mayor valor indicado en el rango [. - .]. El parámetro d representa el exponente en la función kernel polinómica. γ es un factor de ponderación de la medida aplicada sobre dos puntos x e y . r corresponde a un desplazamiento en el espacio. C es el parámetro que controla las variables de holgura.

2.7.5) Árboles de decisión.

Un árbol de decisión corresponde a un clasificador que va particionando linealmente el espacio de variables a medida que se acotan los valores tomados por cada una para un modelo dado. Por ejemplo, un profesor puede utilizar un árbol de decisión para clasificar a sus alumnos en eximibles o no eximibles (Fig 21). Así, un alumno con

$$\text{PROMEDIO} > 5.5$$

es eximible, mientras que un alumno con

$$\text{PROMEDIO} < 5.5 \text{ Y } \text{PROMEDIO} > 5.0 \text{ Y } \text{NOTAS_PARCIALES} < 4.0$$

no es eximible.

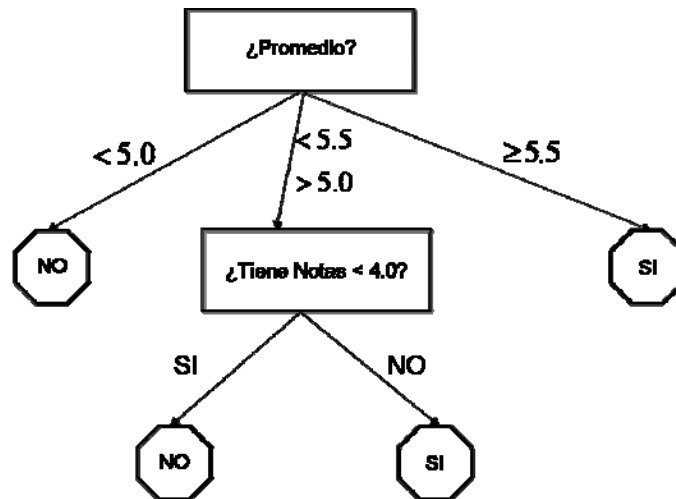


Fig 21. Ejemplo de árbol de decisión. En este ejemplo, se define un criterio para la eximición de alumnos en un curso dado en base a dos variables: el promedio en el curso y las notas parciales en cada evaluación.

Si bien los árboles de decisión son excelentes herramientas de clasificación en lo que se refiere a la fácil comprensión de los resultados, el tiempo de generación de un árbol crece exponencialmente con el número de variables a evaluar, pues se deben probar todas las combinaciones posibles de criterios de partición del espacio de variables. Es por esto que se ha decidido utilizar el algoritmo genético GA Logic, descrito más abajo, ya que permite aprovechar las ventajas de un algoritmo genético en cuanto a la búsqueda paralela de soluciones (debido a los múltiples individuos que componen a la población) de manera de generar en un tiempo razonable árboles que son buenas soluciones al problema de clasificación.

2.7.6) Algoritmos genéticos

Un algoritmo genético [36, 88] es un algoritmo de aprendizaje que considera el proceso de selección natural para alcanzar el máximo fitness (o mínimo costo) sobre una población de individuos. La selección natural corresponde a un proceso evolutivo en el cual los fenotipos – características visibles – que favorecen la adaptación al entorno de los individuos de una población, son seleccionados. Los pasos que definen a un algoritmo genético son los siguientes (Fig 22):

1) Inicialización de la población

El algoritmo se inicia con la generación de una **población**. Una población corresponde a un conjunto de cromosomas (cada cromosoma puede entenderse como un conjunto de genes) que codifican las variables a optimizar. La idea detrás de esto es generar una población de potenciales soluciones en el espacio de búsqueda.

2) Evaluación de una función de fitness sobre cada individuo

Una vez generada la población, es necesario evaluar para cada individuo su “adaptación al medio”. Esto se mide a través de una **función de fitness**. Aquellos individuos con un mayor fitness tienen una mayor probabilidad de sobrevivir, mientras que aquellos con menor fitness tienden a desaparecer de la población.

3) Aplicación de la selección natural

Una vez conocida la adaptación de cada individuo al medio, se aplica un operador de selección de manera de favorecer a aquellos con mejor fitness.

4) Generación de descendencia aplicando variaciones aleatorias

Con el fin de generar nuevas soluciones en el espacio de búsqueda, se generan descendientes de los miembros de la población, para lo cual se aplican distintos operadores que simulan variaciones aleatorias biológicas sobre los cromosomas.

Los operadores más utilizados son aquellos de **mutación**, y **cruce** (o crossing over). El operador de mutación realiza cambios aleatorios de uno o más valores codificados en los cromosomas, lo que permite aumentar la diversidad genética en la población. El operador de cruce intercambia material genético entre cromosomas. Estos operadores son estocásticos y se aplican con probabilidades definidas por el usuario; por lo general la probabilidad de mutación es pequeña e inferior a aquella de cruce.

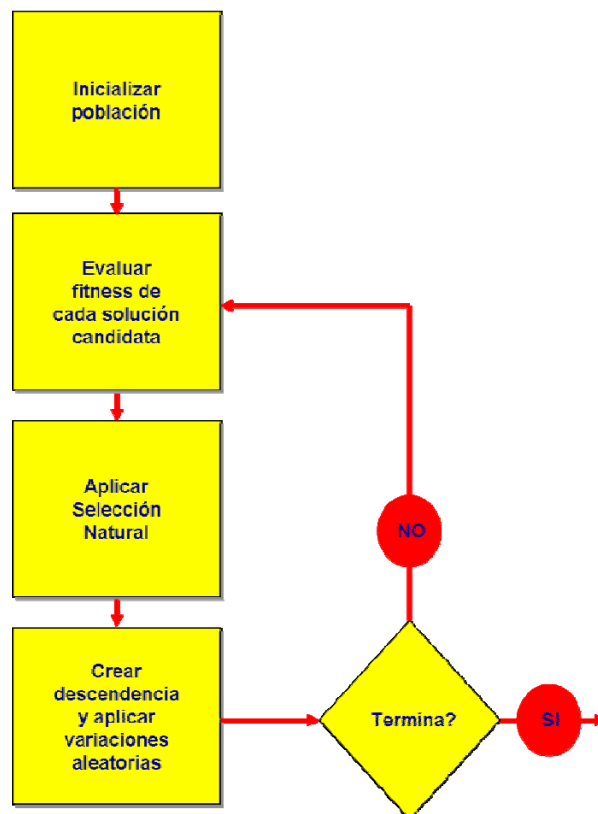


Fig 22. Diagrama de flujo de un algoritmo genético general. Un conjunto de cromosomas inicial que codifica el problema de optimización representa la población inicial. Luego, el ajuste o fitness de cada solución posible al problema es evaluado, y se aplica selección natural de manera de preservar aquellos individuos con mejor fitness. Nueva descendencia es creada con operadores que introducen variaciones aleatorias. Una vez que se alcanza cierto criterio, se detiene el proceso de optimización.

En este trabajo, se utilizaron dos algoritmos genéticos, que se describen a continuación.

2.7.6.1) GA logic

Este algoritmo genético evoluciona las ramas de un árbol de decisión, de manera de generar un clasificador (correspondiente a un árbol de decisión) que se utilice para clasificar los modelos en correctos e incorrectos. Cada cromosoma está compuesto por un número n de genes. Cada gen a su vez está compuesto por: (i) un campo denominado variable, que define el atributo a utilizar, (ii) un campo denominado valor, que define el valor contra el cual se compara el atributo, (iii) un operador de comparación, que para el caso de variables continuas corresponde a los casos $>$ y \leq , y (iv) un operador lógico, que corresponde a los casos AND y OR. La función de fitness que se utiliza en este caso corresponde a la precisión (accuracy) del árbol generado en cada iteración, esto es, al número de modelos en el conjunto de entrenamiento correctamente clasificados, sobre el total de modelos en ese conjunto. Los parámetros utilizados se muestran en la Tabla 4.

Tabla 4. Parámetros utilizados en la ejecución de GA Logic.

PARAMETRO	VALOR
Número de Cromosomas	500
Número Iteraciones	100
Tamaño del Elitismo	1
Tasa de Cruce	0.8
Tasa de Mutación	0.05

Tasa de cruce: Proporción de veces en las que se debe llevar a cabo el cruce. Tamaño del elitismo: cantidad de individuos que pasan directamente a la próxima generación. Pasan los mejores. Número de Iteraciones: cantidad de iteraciones a ejecutar. Tasa de mutación: proporción de veces en la que se debe llevar a cabo la mutación. Número de Cromosomas: cantidad de individuos que componen una generación.

2.7.6.2) GA math

Este algoritmo genético evoluciona fórmulas matemáticas que sirven de clasificadores sobre los modelos. Cada cromosoma está compuesto por un número n de genes. Cada gen a su vez está compuesto por un campo denominado tipo, que define si el individuo será un operando o un operador matemático, y un campo denominado valor, que define el valor que toma el gen en caso de ser un atributo, o el tipo de operación que realiza en caso de ser un operador. La función de fitness utilizada en este caso corresponde a $VP \cdot VN$, donde VP (Verdaderos Positivos) son aquellos modelos correctos clasificados como correctos, y VN (Verdaderos Negativos) son aquellos modelos incorrectos clasificados como incorrectos. Los parámetros utilizados se muestran en la Tabla 5.

Tabla 5. Parámetros utilizados en la ejecución de GA Math.

PARAMETRO	VALOR
Tamaño de la población	300
Máx Iteraciones	100
Tamaño del Elitismo	2
Tasa de Cruce	1
Tasa de Mutación	0.1
Probabilidad "Leave-at-root"	0.2
Máx Normalización	0.8

Tasa de cruce: Proporción de veces en las que se debe llevar a cabo el cruce. Si en una pareja de padres no se ejecuta el cruce, ambos pasan directamente a la próxima generación. Tamaño del elitismo: cantidad de individuos que pasan directamente a la próxima generación. Pasan los mejores. Probabilidad "Leave-at-root": probabilidad de que la raíz de una función matemática sea una variable. Max Iteraciones: cantidad de iteraciones a ejecutar. Tasa de mutación: proporción de veces en la que se debe llevar a cabo la mutación. Max Normalización: parámetro del sistema de normalización lineal. Tamaño de la población: cantidad de individuos que componen una generación.

2.7.7) Perceptrón multicapa

Las redes neuronales artificiales [94] corresponden a un algoritmo de aprendizaje cuyos fundamentos se basan en el modelamiento de las redes neuronales biológicas.

Una neurona biológica funciona de la siguiente manera:

- Las neuronas recogen la información en forma de impulsos proveniente de otras neuronas o receptores que están en las células sensoriales de distintas partes del organismo.
- La integran en un código de activación propio de la célula.
- La transmiten codificada en forma de impulsos a través de su axón.
- El axón distribuye la información en forma de impulsos a través de sus ramificaciones.
- Los impulsos pasan a otras neuronas o a los efectores, quienes reciben la información y la interpretan.

Uno de los primeros modelos artificiales de neuronas fue el **perceptrón** [95, 96] (Fig 23).

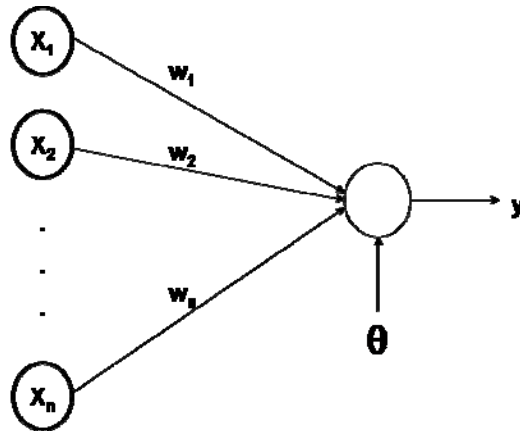


Fig 23. Arquitectura de un perceptrón. El perceptrón recibe n entradas, que se combinan linealmente con un peso w_n asociado a cada una de ellas. Una función de la combinación lineal de las entradas se compara con un umbral de activación θ del perceptrón, lo que define su valor de salida y .

Las entradas están dadas por (x_1, \dots, x_n) , la salida por y , los pesos de las entradas por (w_1, \dots, w_n) y el umbral que se utiliza para decidir la activación de la neurona (el envío de información, dado por y) está dado por θ . La salida de la neurona se calcula como

$$y = F\left(\sum_{i=1}^n w_i x_i + \theta\right)$$

con $F(s) = 1$ si $s > 0$, y $F(s) = -1$ si $s \leq 0$.

Así, esta función actúa como un clasificador binario que asigna el valor 1 a los modelos correctos y el valor -1 a los modelos incorrectos.

El aprendizaje de la neurona ocurre gracias a un conjunto de entrenamiento R , de manera que se deben obtener valores de (w_1, \dots, w_n) asociados al hiperplano

$$w_1 x_1 + w_2 x_2 + \dots + w_n x_n + \theta = 0$$

tal que para los modelos de una clase

$$w_1 x_1 + w_2 x_2 + \dots + w_n x_n + \theta > 0$$

y para los modelos de otra clase

$$w_1 x_1 + w_2 x_2 + \dots + w_n x_n + \theta < 0$$

La obtención de los pesos que logran este hiperplano separador es un proceso iterativo en el cual los pesos se van ajustando gradualmente, produciéndose el **aprendizaje** de la neurona.

El perceptrón descrito sólo genera clasificadores lineales, por lo que su aplicabilidad es limitada. En 1969, Minsky y Papert [97] sugirieron que la combinación de perceptrones podía ser una solución al problema de la separabilidad no lineal, pero no fue hasta 1986 en que Rumelhart, Hinton y Williams [98] propusieron una forma de aprendizaje eficiente (llamada retropropagación) de esta combinación de perceptrones, expandiendo la aplicabilidad de las redes neuronales a problemas no linealmente separables.

El **perceptrón multicapa**, que es una generalización del perceptrón simple, tiene sus neuronas agrupadas en 3 capas distintas (Fig 24):

- Una capa de entrada, cuya única función es recibir las señales de entrada y distribuirlas a la siguiente capa.
- Una o más capas ocultas, que realizan un procesamiento no lineal de los patrones recibidos.
- Una capa de salida, que proporciona la respuesta de la red a las señales de entrada.

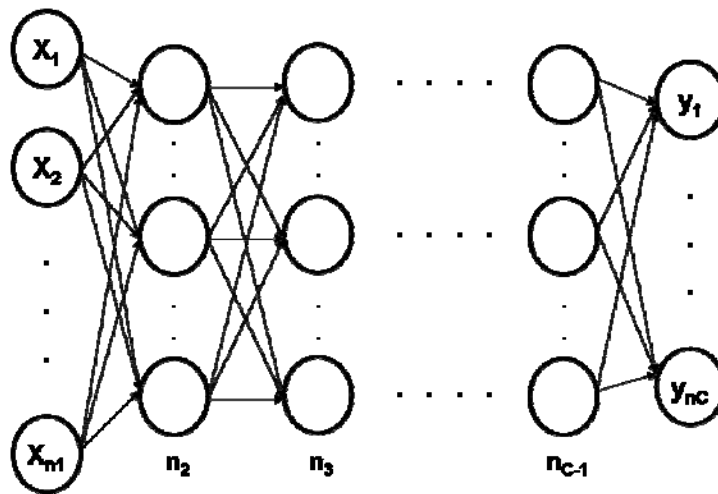


Fig 24. Arquitectura de un perceptrón multicapa. La arquitectura del perceptrón multicapa consta de tres tipos de capas: una capa de entrada, una o más capas ocultas, y una capa de salida. Esto proporciona una estructura que permite encontrar soluciones a problemas de clasificación que no son linealmente separables.

A continuación se formaliza el perceptrón multicapa, de manera de plantear el problema de optimización asociado a este algoritmo de aprendizaje.

Sea un perceptrón multicapa con C capas, de las cuales $C - 2$ son capas ocultas, y n_k neuronas en la capa $k = 1, 2, \dots, C$. Sea $W^k = (w_{ij}^k)$ la matriz de pesos asociados a las conexiones de la capa k a la capa $k + 1$ para $k = 1, 2, \dots, C - 1$, donde w_{ij}^k representa el peso de la conexión de la neurona i de la capa k a la neurona j de la capa $k + 1$.

Sea $\mathbf{U}^k = (\mathbf{u}_i^k)$ el vector de umbrales de las neuronas de la capa \mathbf{k} para $\mathbf{k} = 2, 3, \dots, \mathbf{C}$. Se denota \mathbf{a}_i^k a la activación de la neurona \mathbf{i} de la capa \mathbf{k} . Estas activaciones se calculan como sigue:

- Activación de las neuronas de la capa de entrada: $\mathbf{a}_i^1 = \mathbf{x}_i$, $\mathbf{i} = 1, \dots, \mathbf{n}_1$.
- Activación de las neuronas de la capa oculta \mathbf{k} : $\mathbf{a}_i^k = \mathbf{f} \left(\sum_{j=1}^{\mathbf{n}_{k-1}} \mathbf{w}_{ji}^{k-1} \mathbf{a}_{ji}^{k-1} + \mathbf{u}_i^k \right)$ para $\mathbf{i} = 1, 2, \dots, \mathbf{n}_k$ y $\mathbf{k} = 2, 3, \dots, \mathbf{C} - 1$.
- Activación de las neuronas de la capa de salida: $\mathbf{y}_i = \mathbf{a}_i^{\mathbf{C}} = \mathbf{f} \left(\sum_{j=1}^{\mathbf{n}_{\mathbf{C}-1}} \mathbf{w}_{ji}^{\mathbf{C}-1} \mathbf{a}_j^{\mathbf{C}-1} + \mathbf{u}_i^{\mathbf{C}} \right)$ para $\mathbf{i} = 1, 2, \dots, \mathbf{n}_{\mathbf{C}}$, donde $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_{\mathbf{n}_{\mathbf{C}}})$ es el vector de salida de la red.

La función \mathbf{f} corresponde a la **función de activación**, y existen dos principales:

Función sigmoideal: $\mathbf{f}_1(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{x}}}$, que toma valores en el intervalo $[0, 1]$.

Función tangente hiperbólica: $\mathbf{f}_2(\mathbf{x}) = \frac{1 - e^{-\mathbf{x}}}{1 + e^{-\mathbf{x}}}$, que toma valores en el intervalo $[-1, 1]$.

Se puede demostrar que $\mathbf{f}_2(\mathbf{x}) = 2 \cdot \mathbf{f}_1(\mathbf{x}) - 1$.

Con esto, se puede plantear el problema de optimización que permite ajustar los pesos del perceptrón multicapa.

Se define la función de error como $\mathbf{E} = \frac{1}{\mathbf{N}} \sum_{\mathbf{s}=1}^{\mathbf{N}} \mathbf{e}(\mathbf{h})$, donde \mathbf{N} es el número de modelos,

$\mathbf{e}(\mathbf{h}) = \frac{1}{2} \sum_{i=1}^{\mathbf{n}_{\mathbf{C}}} (\mathbf{s}_i(\mathbf{h}) - \mathbf{y}_i(\mathbf{h}))^2$, $\mathbf{y}(\mathbf{h}) = (\mathbf{y}_1(\mathbf{h}), \dots, \mathbf{y}_{\mathbf{n}_{\mathbf{C}}}(\mathbf{h}))$ es el vector de salidas de la red para el modelo \mathbf{h} y $\mathbf{s}(\mathbf{h}) = (\mathbf{s}_1(\mathbf{h}), \dots, \mathbf{s}_{\mathbf{n}_{\mathbf{C}}}(\mathbf{h}))$ es el vector de salidas deseadas para el modelo \mathbf{h} .

El problema de optimización es

$$\underset{\mathbf{w}}{\text{MIN}}(\mathbf{E})$$

esto es, encontrar aquella matriz de pesos \mathbf{W}^* que minimiza el error, de manera que las salidas arrojadas por la red para cada modelo sean tan cercanas como sea posible a las salidas deseadas.

El método más utilizado para la búsqueda de la solución óptima al problema de minimización planteado es aquél de **descenso por el gradiente**. El descenso por el gradiente se logra planteando, para cada parámetro \mathbf{w} :

$$\mathbf{w}(\mathbf{h}) = \mathbf{w}(\mathbf{h} - 1) - \alpha \frac{\partial \mathbf{e}(\mathbf{h})}{\partial \mathbf{w}} + \eta \Delta \mathbf{w}(\mathbf{h} - 1)$$

Con $\Delta \mathbf{w}(\mathbf{n} - 1) = \mathbf{w}(\mathbf{n} - 1) - \mathbf{w}(\mathbf{n} - 2)$.

El parámetro α es la **razón de aprendizaje**, y determina la magnitud de los pasos dados sobre la superficie de error para ajustar el peso \mathbf{w} .

El parámetro η es el **momento**, y sirve para compensar el efecto negativo de valores muy grandes de α , para los cuales el algoritmo puede mantenerse oscilante en torno al mínimo o simplemente saltarlo; a su vez η sirve para compensar el efecto negativo de valores pequeños de α , para los cuales la velocidad de convergencia se torna muy lento.

En 1986, Rumelhart [98] propuso un método eficiente para aplicar el método de descenso por el gradiente sobre las distintas capas del perceptrón multicapa, y que se conoce como **algoritmo de retropropagación**. En palabras, este algoritmo plantea que [94]:

Cada neurona de salida distribuye hacia atrás su error a todas las neuronas ocultas que se conectan a ella, ponderado por el valor de la conexión. De este modo, cada neurona oculta recibe un cierto error de cada neurona de salida, y la suma de estas cantidades es el error asociado a la neurona oculta. Dichos valores permiten a su vez obtener los errores de las capas que la anteceden, y así sucesivamente hasta llegar a la primera capa oculta. De ahí viene el nombre de **algoritmo de retropropagación**, pues los errores para las neuronas de la red se retropropagan hacia todas las neuronas de la capa oculta.

Dado que ya tenemos una forma eficiente de ajustar los pesos de la red neuronal, podemos establecer el proceso de aprendizaje del perceptrón multicapa como sigue:

- 1) Inicializar los pesos y umbrales de la red con valores aleatorios.
- 2) Tomar un modelo \mathbf{h} del conjunto de entrenamiento, y propagar hacia la salida los valores que este modelo toma para cada variable, dados por $\mathbf{x}(\mathbf{h})$. Obtener el vector de salida de la red, $\mathbf{y}(\mathbf{h})$.
- 3) Evaluar el error $\mathbf{e}(\mathbf{h})$ sobre el modelo \mathbf{h} .
- 4) Aplicar el algoritmo de retropropagación para modificar los pesos y umbrales de la red.
- 5) Repetir los pasos 2)-4) para todos los modelos en el conjunto de entrenamiento. Esto define un **ciclo de aprendizaje**.
- 6) Evaluar el error total, \mathbf{E} .
- 7) Repetir los pasos 2)-6) un número \mathbf{m} de ciclos de aprendizaje, o hasta que el error haya alcanzado un mínimo.

Con el fin de evitar un sobreajuste a los datos de entrenamiento de la solución que entrega la red, se debe utilizar un conjunto adicional de modelos independientes que actúen como control en cada iteración (Fig 25). Así, se puede contar un número de ciclos de aprendizaje consecutivos en los que el error sobre el conjunto de control crece, tras los cuales se debe detener el proceso de aprendizaje.

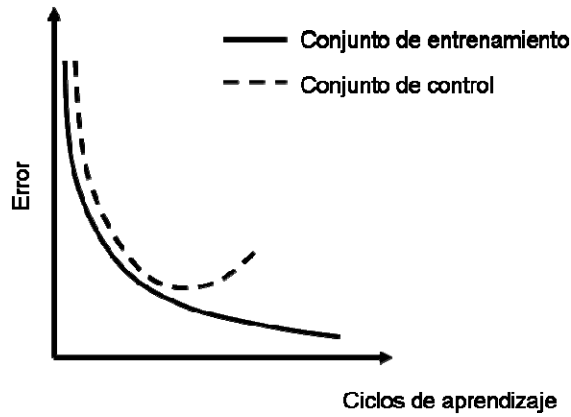


Fig 25. Control del sobreajuste en un perceptrón multicapa. Para cada ciclo de aprendizaje, el error se mide tanto sobre el conjunto de entrenamiento como sobre el conjunto de control, de manera de detener el ajuste de los pesos del perceptrón multicapa cuando el error sobre el conjunto de control crece de manera consecutiva por 15 ciclos.

El espacio de valores posibles y muestreados para los parámetros del MLP se muestran en Tabla 6.

Tabla 6. Espacio de valores posibles y muestreados para los parámetros de MLP.

Parámetro	Valores Posibles	Valores Muestreados
Número de Capas Ocultas	\mathbb{N}	1,2
Número de Neuronas primera capa oculta	\mathbb{N}	1,2,4,8,12,16,20,22
Número de Neuronas segunda capa oculta	\mathbb{N}	1,2,4,8,12,16,20,22
Tasa de aprendizaje	(0-1)	[0.1-0.4] p:0.1
Tasa de Momentum	[0-1]	[0.0-0.3] p:0.1

La notación p: 0.1 indica que la variable en cuestión se varió con un paso de 0,1 desde el menor al mayor valor indicado en el rango [. - .].

2.7.8) Redes Bayesianas

Las redes bayesianas corresponden a un algoritmo de aprendizaje con una estructura de grafo acíclico dirigido, en el cual cada nodo representa una variable y la relación entre variables se estima mediante la probabilidad condicional de un nodo hijo dado sus padres.

Sea $\mathbf{U} = \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_n\}$ un conjunto de variables. Una **red Bayesiana B** sobre U corresponde a $\mathbf{B} = (\mathbf{B}_s, \mathbf{B}_p)$, donde \mathbf{B}_s es la estructura de la red, correspondiente a un grafo acíclico dirigido sobre las variables en U, y $\mathbf{B}_p = \{p(\mathbf{u} | \mathbf{pa}(\mathbf{u})) : \mathbf{u} \in \mathbf{U}\}$ corresponde a un conjunto de tablas de probabilidad para cada variable, donde $\mathbf{pa}(\mathbf{u})$ corresponde a los padres de \mathbf{u} en \mathbf{B}_s .

El problema de aprendizaje en redes bayesianas es el siguiente: dado un conjunto de datos D sobre U, encontrar el grafo dirigido acíclico \mathbf{B}_s que mejor represente el conjunto de dependencias (e independencias) entre los datos.

Sea $\mathbf{x}_0 = \mathbf{y}$ la clase respuesta, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ las variables o atributos, e imponemos $\mathbf{x}_0 = \mathbf{y}$ como el nodo padre de toda la red, entonces para usar la red bayesiana como clasificador basta calcular

$$\arg \max_{\mathbf{y}} \mathbf{P}(\mathbf{y} | \mathbf{x})$$

donde

$$\mathbf{P}(\mathbf{y} | \mathbf{x}) = \frac{\mathbf{P}(\mathbf{U})}{\mathbf{P}(\mathbf{x})} \propto \prod_{i=0}^n \mathbf{P}(\mathbf{x}_i | \mathbf{pa}(\mathbf{x}_i))$$

Esta forma de escribir la probabilidad es similar a aquella para naive Bayes, en cuyo caso $\mathbf{pa}(\mathbf{x}_i)$ corresponde a la clase respuesta, $\forall i$. Así, el clasificador naive Bayes es un caso particular de las redes Bayesianas utilizadas como clasificadores, y la generalización radica en que estas últimas van un paso más allá en el modelamiento de las dependencias utilizando una relación del tipo

$$\mathbf{I}(\mathbf{R}, \mathbf{S} | \mathbf{T}) \Leftrightarrow \mathbf{P}(\mathbf{R} | \mathbf{S}, \mathbf{T}) = \mathbf{P}(\mathbf{R} | \mathbf{T})$$

donde $\mathbf{I}(\cdot, \cdot | \cdot)$ representa independencia condicional, y $\mathbf{R}, \mathbf{S}, \mathbf{T}$ son cualquier subconjunto de \mathbf{U} .

Para encontrar una red bayesiana óptima dentro de aquellas posibles, es necesario contar con lo siguiente:

- una métrica de la adaptación de los datos \mathbf{D} a una estructura \mathbf{B}_s ,
- un algoritmo de búsqueda que permita encontrar la solución que maximiza esa métrica sobre las estructuras posibles, y
- una manera de estimar las probabilidades condicionales $\mathbf{p}(\mathbf{x}_i | \mathbf{pa}(\mathbf{x}_i))$.

2.7.8.1) Métrica de adaptación de \mathbf{D} a \mathbf{B}_s .

Sea r_i , $1 \leq i \leq n$ la cardinalidad de la variable \mathbf{x}_i , suponiendo que es discreta.

Sea q_i la cardinalidad del conjunto de padres de \mathbf{x}_i en \mathbf{B}_s , esto es, $q_i = \prod_{\mathbf{x}_j \in \mathbf{pa}(\mathbf{x}_i)} r_j$. Si

$\mathbf{pa}(\mathbf{x}_i) = \Phi$, entonces $q_i = 1$.

Sea N_{ij} , $1 \leq i \leq n$, $1 \leq j \leq q_i$ el número de instancias en \mathbf{D} para los cuales $\mathbf{pa}(\mathbf{x}_i)$ toma su j -ésimo valor.

Sea N_{ijk} , $1 \leq i \leq n$, $1 \leq j \leq q_i$, $1 \leq k \leq r_i$ el número de instancias en \mathbf{D} para los cuales $\mathbf{pa}(\mathbf{x}_i)$ toma su j -ésimo valor y para los cuales \mathbf{x}_i toma su k -ésimo valor. Así,

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk}.$$

Sea \mathbf{N} el número de instancias en \mathbf{D} .

El problema se trata de calcular $\mathbf{P}(\mathbf{B}_s | \mathbf{D})$. Aplicando la regla de Bayes,

$$\mathbf{P}(\mathbf{B}_s | \mathbf{D}) = \frac{\mathbf{P}(\mathbf{D} | \mathbf{B}_s) \cdot \mathbf{P}(\mathbf{B}_s)}{\mathbf{P}(\mathbf{D})}$$

Para efectos de buscar aquella red que maximiza esta métrica por definir, se asume $\mathbf{P}(\mathbf{B}_s)$ uniforme, y ya que $\mathbf{P}(\mathbf{D})$ es el mismo valor para todos los casos, el único término que hace la diferencia es

$$\mathbf{P}(\mathbf{B}_s | \mathbf{D}) \propto \mathbf{P}(\mathbf{D} | \mathbf{B}_s)$$

La **métrica bayesiana** para una estructura \mathbf{B}_s dado un conjunto de datos \mathbf{D} está dada por

$$\mathbf{P}(\mathbf{B}_s | \mathbf{D}) \propto \mathbf{P}(\mathbf{D} | \mathbf{B}_s) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\mathbf{N}'_{ij})}{\Gamma(\mathbf{N}'_{ij} + \mathbf{N}_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\mathbf{N}'_{ijk} + \mathbf{N}_{ijk})}{\Gamma(\mathbf{N}'_{ijk})}$$

Donde $\Gamma(\mathbf{z}) = \int_0^{\infty} \mathbf{t}^{\mathbf{z}-1} \mathbf{e}^{-\mathbf{t}} \mathbf{d}\mathbf{t}$ es la función gamma, \mathbf{N}'_{ij} y \mathbf{N}'_{ijk} son parámetros tal que

$$\mathbf{N}'_{ij} = \sum_{k=1}^{r_i} \mathbf{N}'_{ijk}$$

2.7.8.2) Algoritmos de búsqueda

Una vez definida la métrica de adaptación de un conjunto de datos \mathbf{D} a una estructura de red \mathbf{B}_s , es necesario definir una forma de búsqueda sobre las estructuras posibles de manera de encontrar en forma eficiente aquella que maximiza el valor de la métrica.

2.7.8.2.1) Algoritmo K2

El algoritmo K2 [99] trabaja sobre el principio de que las variables están ordenadas, de manera que los posibles padres de una variable aparecen en el orden antes que ella misma. Esto acota el espacio de búsqueda de los posibles padres para cada variable.

El algoritmo se inicia con el conjunto de padres para cada variable siendo el conjunto vacío. Luego, para cada variable, se compara la medida aplicada sobre los padres actuales de cada variable y aquella obtenida al introducir una variable predecesora al conjunto de padres. Aquella que sea mayor definirá el nuevo conjunto de padres de la respectiva variable.

2.7.8.2.2) Algoritmo hill climbing

El algoritmo de hill climbing [100] se basa en la idea de ir ascendiendo por el gradiente de la métrica de adaptación de los datos a la estructura, utilizando alguna definición de vecindad.

El algoritmo parte de una solución inicial, que puede ser por ejemplo una estructura propia de naive Bayes. A partir de esta solución se calcula el nuevo valor de la métrica utilizada sobre todos los grafos vecinos a la solución actual, tomando como nueva solución aquella que maximice la métrica.

Los grafos vecinos se definen como cualquier estructura \mathbf{B}'_s resultante de incluir un solo arco, eliminar un solo arco o invertir el sentido de un arco existente en \mathbf{B}_s , cuidando de formar ciclos.

El algoritmo se detiene cuando no existe ningún vecino que pueda mejorar la situación actual, esto es, cuando se alcanza un máximo local.

2.7.8.2.3) TAN

El algoritmo TAN (Tree Augmented Naive Bayes) [101, 102] restringe la topología de la red a un árbol. Para realizar el aprendizaje, este algoritmo se basa en el concepto de información mutua:

$$\mathbf{I}(\mathbf{x}, \mathbf{Y} | \mathbf{C}) = \sum_{\mathbf{x}, \mathbf{y}, \mathbf{c}} \mathbf{P}(\mathbf{x}, \mathbf{y}, \mathbf{c}) \log \left(\frac{\mathbf{P}(\mathbf{x}, \mathbf{y}, \mathbf{c})}{\mathbf{P}(\mathbf{x} | \mathbf{c}) \cdot \mathbf{P}(\mathbf{y} | \mathbf{c})} \right)$$

Luego, el algoritmo está compuesto de las siguientes etapas:

- Calcula $\mathbf{I}(\mathbf{x}_i, \mathbf{x}_j | \mathbf{C})$ para todo par de atributos $i \neq j$.
- Crea un grafo no dirigido con todos los atributos como conjunto de nodos y añade aristas entre cada par de nodos.
- Asocia a cada arista el peso $\mathbf{I}(\mathbf{x}_i, \mathbf{x}_j | \mathbf{C})$.
- Construye un árbol expandido de máximo peso a partir del grafo anterior, utilizando el algoritmo de Chow y Liu [103].
- Elige un nodo cualquiera del árbol como raíz y direcciona a partir de él el resto de las aristas.
- Añade la variable respuesta \mathbf{C} y el conjunto de aristas dirigidas $\mathbf{C} \rightarrow \mathbf{x}_i$ para todo atributo \mathbf{x}_i .
- Arroja la red obtenida.

2.7.8.2.4) Annealing simulado

Este algoritmo de búsqueda genera al azar una estructura candidata \mathbf{B}'_s cercana a la actual, \mathbf{B}_s . Si $\mathbf{P}(\mathbf{B}_s | \mathbf{D}) < \mathbf{P}(\mathbf{B}'_s | \mathbf{D})$, se acepta la estructura \mathbf{B}'_s . Si $\mathbf{P}(\mathbf{B}_s | \mathbf{D}) > \mathbf{P}(\mathbf{B}'_s | \mathbf{D})$, se acepta la estructura \mathbf{B}'_s con probabilidad

$$\exp(\tau_i (\mathbb{P}(\mathbf{B}_s' | \mathbf{D}) - \mathbb{P}(\mathbf{B}_s | \mathbf{D})))$$

donde τ_i es la temperatura para la iteración i -ésima. La temperatura parte en un valor τ_0 y decrece de manera que $\tau_{i+1} = \delta \cdot \tau_i$. El algoritmo se aplica por un número predefinido de pasos, hasta que se detiene.

2.7.8.3) Estimación de las tablas de probabilidad.

Para estimar la tabla de probabilidad asociada a cada nodo, se utiliza la estimación directa, como sigue:

$$\mathbb{P}(\mathbf{x}_i = \mathbf{k} | \mathbf{pa}(\mathbf{x}_i) = \mathbf{j}) = \frac{\mathbf{N}_{i,jk} + \mathbf{N}'_{i,jk}}{\mathbf{N}_{i,j} + \mathbf{N}'_{i,j}}$$

2.8) Uso de la base de datos para la obtención de clasificadores.

2.8.1) Conjuntos de entrenamiento, validación y prueba.

Utilizar de manera correcta la base de datos para entrenar a los algoritmos de aprendizaje y poner a prueba los clasificadores generados es clave cuando se busca comparar distintos métodos de clasificación [104].

La primera aproximación que se puede considerar es utilizar la base de datos completa tanto para entrenar a los algoritmos de aprendizaje como para poner a prueba el rendimiento de los distintos clasificadores. Sin embargo, proceder de esta forma es erróneo pues resulta en sobrestimaciones de la medida de rendimiento del clasificador. Así, aquél clasificador que más se sobreajuste a los datos, tendrá un mejor rendimiento y será elegido como clasificador óptimo, lo cual es una mala solución pues el objetivo es obtener un clasificador que represente una solución lo más genérica posible en términos de separar los modelos en correctos e incorrectos.

Una segunda aproximación es separar la base de datos en dos conjuntos disjuntos: un **conjunto de entrenamiento R**, que se utiliza para entrenar los algoritmos de aprendizaje, y un **conjunto de prueba T**, que se utiliza para medir el rendimiento de cada clasificador. La generación de cada conjunto a partir de la base de datos inicial debe ser al azar con el fin de no inducir sesgo alguno.

Una tercera aproximación que se debe considerar cuando los algoritmos de aprendizaje tienen parámetros que deben ser optimizados, corresponde a la generación de un tercer conjunto adicional a aquellos de entrenamiento y prueba, y que corresponde a un **conjunto de validación V**. Así, un algoritmo de aprendizaje es entrenado con **R**, sus parámetros son optimizados con **V**, y el rendimiento del clasificador generado con los parámetros optimizados se mide con el conjunto **T**.

2.8.2) Modalidades de entrenamiento, validación y prueba.

Para aquellos casos en los que la optimización de parámetros no es necesaria, basta con entrenar sobre el conjunto de entrenamiento R y medir el rendimiento de los clasificadores con el conjunto de prueba T.

En los casos en los que se requiere optimizar los parámetros del algoritmo de aprendizaje, es necesario evitar el posible sesgo proveniente de la generación al azar de los conjuntos de entrenamiento y validación. Para esto, se utiliza la validación cruzada de orden k (o k -fold cross validation), la cual consiste en separar el conjunto de entrenamiento en k subconjuntos independientes, de manera de entrenar utilizando $k-1$ subconjuntos, y validar con el subconjunto restante. Esto se repite k veces para cada uno de los subconjuntos generados (Fig 26).

Luego, el conjunto de parámetros óptimo del algoritmo de aprendizaje será aquél que hace máximo el rendimiento promedio sobre las k instancias de entrenamiento y validación.

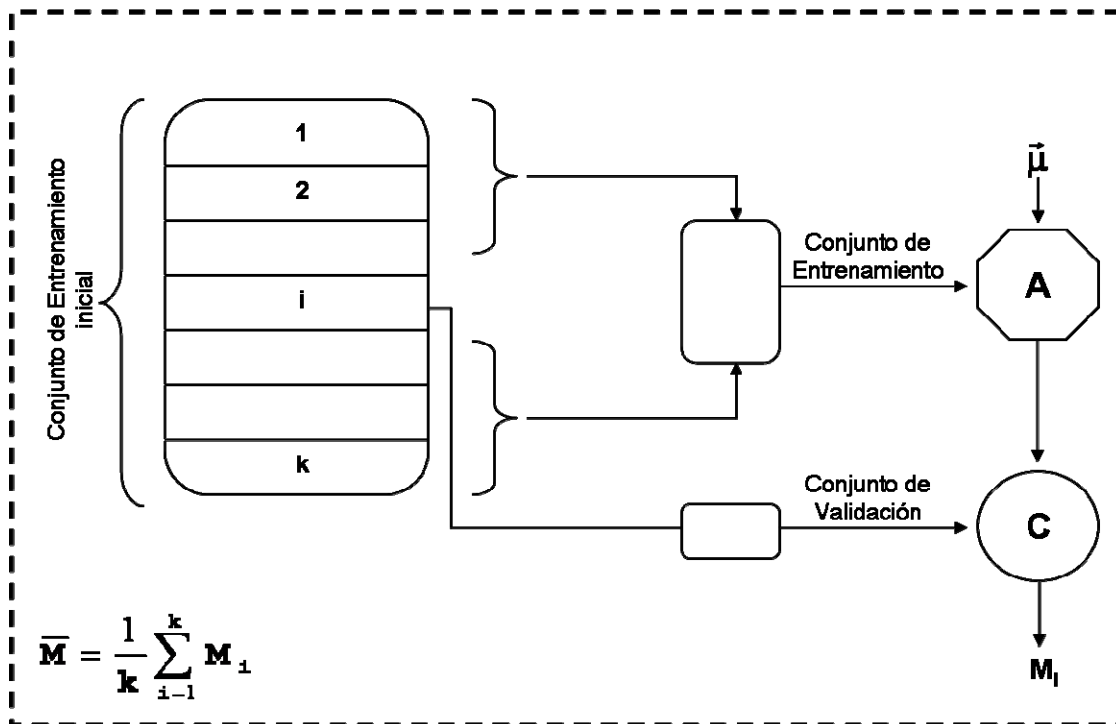


Fig 26. Validación cruzada de orden k . A: algoritmo de aprendizaje. C: clasificador M_i : rendimiento para la iteración i -ésima. $\bar{\mu}_i$: instancia del vector de parámetros de A. \bar{M} : rendimiento promedio para una instancia de $\bar{\mu}_i$. El algoritmo de aprendizaje A se entrena k veces con cada valor definido para el vector de parámetros $\bar{\mu}_i$, con todos excepto el k -ésimo subconjunto, el cual se utiliza para evaluar el clasificador generado por A en esa instancia de $\bar{\mu}_i$. Luego, el promedio del rendimiento de los clasificadores generados se calcula. Se elige aquél valor de $\bar{\mu}_i$ que maximiza el rendimiento promedio \bar{M} .

2.9) Medidas de rendimiento de clasificadores.

Al clasificar cada modelo, éste puede ser categorizado en cuatro clases:

Verdadero Positivo (VP): el modelo es correcto y se clasifica como correcto.

Verdadero Negativo (VN): el modelo es incorrecto y se clasifica como incorrecto.

Falso Positivo (FP): el modelo es incorrecto y se clasifica como correcto.

Falso Negativo (FN): el modelo es correcto y se clasifica como incorrecto.

En base a esto, se define la **precisión (o accuracy) del clasificador \hat{f}** como:

$$\text{ACC}(\hat{f}) = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

Esto es, la precisión corresponde a la fracción del total de modelos que son correctamente clasificados.

El hecho de que la precisión sea una buena medida del rendimiento de un clasificador es discutible, depende del contexto, y de lo que significa un FP y un FN.

Sea por ejemplo el caso de un cardiólogo, y una enfermedad coronaria que se da en 1 de cada 100 personas. El rendimiento de este cardiólogo es evaluado anualmente en el hospital en base a la fracción de diagnósticos correctamente realizados sobre esta enfermedad coronaria.

Si el cardiólogo decide clasificar a todos sus pacientes como sanos, entonces tendrá en promedio una precisión de 99%, pues de cada 100 pacientes que llegan a atenderse con él, 99 debieran estar sanos y 1 enfermo. Así, el cardiólogo es considerado un médico casi óptimo.

Pero el costo de clasificar a un paciente como sano, siendo que está enfermo (esto es, un falso negativo) es muy alto desde muchas perspectivas, comenzando con la vida del paciente. Este ejemplo muestra que la precisión como medida del rendimiento de un clasificador es dependiente del contexto y del costo asociado a cometer falsos positivos y falsos negativos, y por lo tanto puede no ser una buena medida si no se consideran esos factores.

La curva ROC (Receiver Operating Characteristic) [105-107] es una herramienta que permite evaluar un clasificador de manera global, entregando en una gráfica bidimensional el comportamiento del clasificador sobre todos los escenarios posibles.

Esta curva se genera variando aquél **umbral** que clasifica a los modelos en correctos e incorrectos, de manera de ir generando distintas tasas de VP y FP.

Se define la **sensibilidad del clasificador \hat{f}** como la proporción de verdaderos positivos.

Se define la **especificidad del clasificador \hat{f}** como la proporción de verdaderos negativos.

En el contexto del cardiólogo, se espera que la clasificación resultante tenga alta sensibilidad y especificidad.

En la curva ROC, lo que se grafica es 1-especificidad (en la abscisa) versus sensibilidad (en la ordenada) (Figura 27). Esto es idéntico a graficar la tasa de falsos positivos (fp) versus la tasa de verdaderos positivos (vp), donde

$$vp = \frac{VP}{VP + FN} ; fp = \frac{FP}{VN + FP}$$

Una medida que resume en un índice el rendimiento del clasificador utilizando la curva ROC es el **área bajo la curva ROC**, o **AUC** (Fig 27).

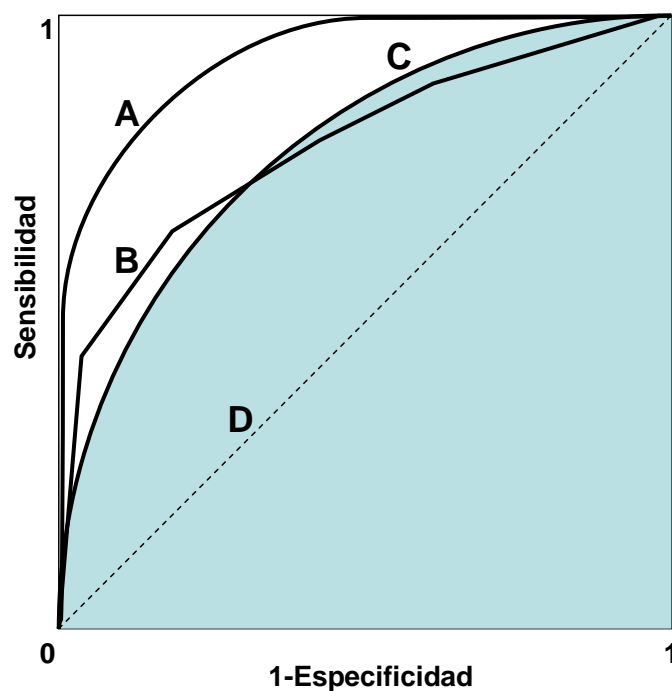


Fig 27. Curvas ROC y AUC. El área sombreada corresponde al AUC para el clasificador C. El clasificador A es aquél de mejor rendimiento, pues tiene un área bajo la curva mayor a los otros clasificadores. El clasificador B es mejor que C sólo en un primer tramo, pero luego el clasificador C supera a C para el tramo siguiente. El clasificador D corresponde al peor caso, en el cual $AUC=0.5$.

Adicionalmente, se ha demostrado que AUC es una medida estadísticamente consistente y con un mayor poder de discriminación que la precisión [108].

2.10) Test estadísticos de comparación de clasificadores

Una vez medido el rendimiento de cada clasificador en términos de su precisión y de su AUC, es necesario medir si la diferencia entre cada par de clasificadores es estadísticamente significativa. Para lograrlo, se recurre a dos tests: el test de McNemar [89, 109] para ser aplicado sobre las tasas de error, y el test de DeLong [110], para ser aplicado sobre el AUC.

2.10.1) Test de McNemar

Dadas las clasificaciones obtenidas sobre el conjunto de prueba por dos clasificadores, $\hat{\mathbf{f}}_1$ y $\hat{\mathbf{f}}_2$, se construye una tabla de contingencia con el fin de contabilizar aquellos casos en que ambos clasificadores aciertan o fracasan en la predicción, como sigue:

n_{00} = Número de modelos mal clasificados tanto por $\hat{\mathbf{f}}_1$ como por $\hat{\mathbf{f}}_2$	n_{01} = Número de modelos mal clasificados por $\hat{\mathbf{f}}_1$ pero bien clasificados por $\hat{\mathbf{f}}_2$
n_{10} = Número de modelos mal clasificados por $\hat{\mathbf{f}}_2$ pero bien clasificados por $\hat{\mathbf{f}}_1$	n_{11} = Número de modelos bien clasificados tanto por $\hat{\mathbf{f}}_1$ como por $\hat{\mathbf{f}}_2$

Con $n = n_{00} + n_{01} + n_{10} + n_{11}$ el número total de modelos en el conjunto de prueba.

La hipótesis nula está dada por

$$H_0 : P(\hat{\mathbf{f}}_1(\mathbf{x}) = \mathbf{f}(\mathbf{x})) = P(\hat{\mathbf{f}}_2(\mathbf{x}) = \mathbf{f}(\mathbf{x}))$$

esto es, ambos clasificadores deben tener la misma tasa de error.

El estadístico que se utiliza para testear esta hipótesis es:

$$Q = \frac{(n_{01} - n_{10} - 1)^2}{n_{01} + n_{10}} \sim \chi_1^2$$

Y la hipótesis nula se rechaza si

$$P(\chi_1^2 > Q) < \alpha$$

Con α usualmente 0.05.

2.10.2) Test de Delong

Sea $\hat{\mathbf{f}}$ un clasificador que calcula sobre cada modelo del conjunto de prueba un valor continuo o score. Para el conjunto de modelos correctos, esos valores son $\mathbf{X}_1, \dots, \mathbf{X}_m$ y para los modelos incorrectos los valores son $\mathbf{Y}_1, \dots, \mathbf{Y}_n$. El área bajo la curva ROC (AUC) se puede estimar utilizando el estadístico de Mann-Whitney:

$$\hat{\theta} = \sum_{j=1}^n \sum_{i=1}^m \psi(\mathbf{x}_i, \mathbf{y}_j)$$

donde

$$\psi(\mathbf{x}_i, \mathbf{y}_j) = \begin{cases} 1 & \mathbf{x}_i > \mathbf{y}_j \\ 0.5 & \mathbf{x}_i = \mathbf{y}_j \\ 0 & \mathbf{x}_i < \mathbf{y}_j \end{cases}$$

Es la función kernel sobre el par $(\mathbf{x}_i, \mathbf{y}_j)$.

Si los modelos correctos tienden a adoptar valores más altos que los valores incorrectos, entonces se utiliza la función kernel tal como se muestra. En el caso que los modelos correctos tienden a tomar valores inferiores que los modelos incorrectos, basta aplicar la función kernel con el criterio de desigualdades invertido.

Dados dos clasificadores $\hat{\mathbf{f}}_1$ y $\hat{\mathbf{f}}_2$ y sus respectivas estimaciones de las AUC, $\hat{\theta}_1$ y $\hat{\theta}_2$ se plantea la hipótesis nula correspondiente a la igualdad de AUC,

$$H_0 : \theta_1 = \theta_2$$

El estadístico que se construye para testear esta hipótesis es

$$Q = \hat{\theta} \cdot \mathbf{L}^t (\mathbf{L} \cdot \mathbf{s} \cdot \mathbf{L}^t) \mathbf{L} \cdot \hat{\theta} \sim \chi_1^2$$

donde $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$, $\mathbf{L} = (1, -1)$ y \mathbf{s} es la matriz de covarianza estimada para $\hat{\theta}$.

La hipótesis nula se rechaza si

$$P(\chi_1^2 > Q) < \alpha$$

Con α usualmente 0.05.

Adicionalmente, se puede construir un intervalo a un $100 \cdot (1 - \alpha)\%$ de nivel confianza utilizando el siguiente estadístico:

$$\frac{\mathbf{L} \cdot \hat{\theta}^t - \mathbf{L} \cdot \theta^t}{\sqrt{(\mathbf{L} \cdot \mathbf{s} \cdot \mathbf{L}^t)}} \sim N(0, 1)$$

donde $\theta = (\theta_1, \theta_2)$.

2.11) Software

Con el fin de contar con herramientas de confianza al momento de llevar a cabo cada uno de los pasos de este trabajo, se buscaron herramientas, cuando fue necesario, que fueran reconocidas por la comunidad científica y que tuvieran un mínimo de documentación que permitiera entender el algoritmo subyacente y el funcionamiento general.

Para la generación de los modelos proteicos, se utilizó el software Modeller 6v0. (<http://salilab.org/modeller>).

Para el cálculo de estadísticos básicos, detección de outliers y análisis de componentes principales, se utilizó el software *StatGraphics Plus 4.0* (<http://www.statgraphics.com>).

Para la generación de los clasificadores de Perceptrón Multicapa, naive Bayes y Redes Bayesianas, se utilizó el software WEKA 3 (<http://www.cs.waikato.ac.nz/ml/weka/>).

Para la generación del clasificador SVM, se utilizó el software LibSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>).

Para la generación de los clasificadores GA Math y GA Logic, se utilizaron implementaciones en C++ y C (resp.) creadas por Prof. Francisco Melo e implementadas por Prof. Francisco Melo y José Tomás Etérovic, este último titulado del programa de magister en Ciencias de la Computación de la Universidad Católica.

Para el resto de los algoritmos de ranking, selección de variables, clasificadores y test de comparación de clasificadores se crearon y utilizaron implementaciones creadas por el autor, en C++.

3) Resultados

En este trabajo se enfrentó el problema de la evaluación de la calidad de un modelo proteico generado computacionalmente a partir de su secuencia primaria de aminoácidos (Fig 28). Para resolver esta tarea, se midieron diferentes atributos o variables correspondientes al alineamiento de la secuencia objetivo con la estructura molde, al modelo generado, y al molde utilizado para generar los modelos (Tabla 7). Esto generó una base de datos correspondiente a 2,299 modelos (1,153 correctos y 1,146 incorrectos) con 31 variables medidas sobre cada uno. Luego, esta base de datos fue analizada con el fin de detectar outliers, eliminar redundancia y maximizar la relevancia de las variables con respecto a la clase respuesta (Fig 29). Finalmente, diferentes algoritmos de aprendizaje fueron utilizados para generar clasificadores, y estos clasificadores fueron evaluados sobre un conjunto de prueba con el fin de definir a uno de ellos como óptimo dentro de aquellos generados (Fig 30).

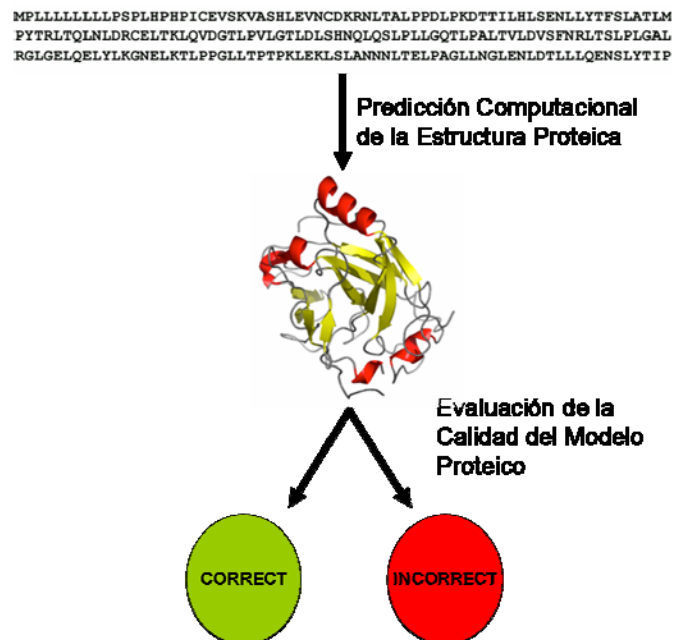


Fig 28. El problema enfrentado en este trabajo. Dada una secuencia primaria de aminoácidos, el modelo proteico generado utilizando modelado comparativo debe evaluarse en términos de su calidad, de manera de clasificarlo en correcto o incorrecto.

Tabla 7. Identificación, descripción y referencia para cada variable utilizada.

	NUM	ID	DESCRIPCION	REF
ALI	1	ALI_SEQIDE	% identidad de secuencia en alineamiento.	[40]
	2	ALI_ZSCO	Z-score del alineamiento.	[40]
	3	ALI_PBEVAL	E-value de PSI-BLAST.	[27]
	4	ALI_TGCOV	Porcentaje de secuencia objetivo modelada.	[42]
MODEL	5	ML_LEN	Largo de la cadena.	N.A.
	6	ML_COMP	Compactación.	[40]
	7	ML_ZCOMB	Z-score de energía combinado.	[42]
	8	ML_ZPAIR	Z-score de energía para pares residuo-residuo.	[42]
	9	ML_ZSURF	Z-score de energía para superficie accesible	[42]
	10	ML_PP	Propensión a la partición.	[53]
	11	ML_COABS	Orden de contacto absoluto.	[54]
	12	ML_COREL	Orden de contacto relativo.	[54]
	13	ML_RG	Radio de giro	[40]
TEMPLATE FRAGMENT	14	TF_LEN	Largo de la cadena.	N.A.
	15	TF_COMP	Compactación.	[40]
	16	TF_ZCOMB	Z-score de energía combinado.	[42]
	17	TF_ZPAIR	Z-score de energía para pares residuo-residuo.	[42]
	18	TF_ZSURF	Z-score de energía para superficie accesible	[42]
	19	TF_PP	Propensión a la partición.	[53]
	20	TF_COABS	Orden de contacto absoluto.	[54]
	21	TF_COREL	Orden de contacto relativo.	[54]
	22	TF_RG	Radio de giro	[40]
	TEMPLATE WHOLE	23	TW_LEN	Largo de la cadena.
24		TW_COMP	Compactación.	[40]
25		TW_ZCOMB	Z-score de energía combinado.	[42]
26		TW_ZPAIR	Z-score de energía para pares residuo-residuo.	[42]
27		TW_ZSURF	Z-score de energía para superficie accesible	[42]
28		TW_PP	Propensión a la partición.	[53]
29		TW_COABS	Orden de contacto absoluto.	[54]
30		TW_CO_REL	Orden de contacto relativo.	[54]
31		TW_RG	Radio de giro	[40]

Para cada variable se define un nombre identificador, una breve descripción y una referencia. ALI: variables medidas sobre el alineamiento secuencia-estructura. MODEL (ML): variables medidas sobre la estructura del modelo generado. TEMPLATE FRAGMENT (TF): variables medidas sobre la estructura de la zona del molde efectivamente utilizada para generar el modelo. TEMPLATE WHOLE (TW): variables medidas sobre la estructura del molde completo utilizado para generar el modelo, sin importar qué fracción de este efectivamente se utilizó. N.A.: No Aplica.

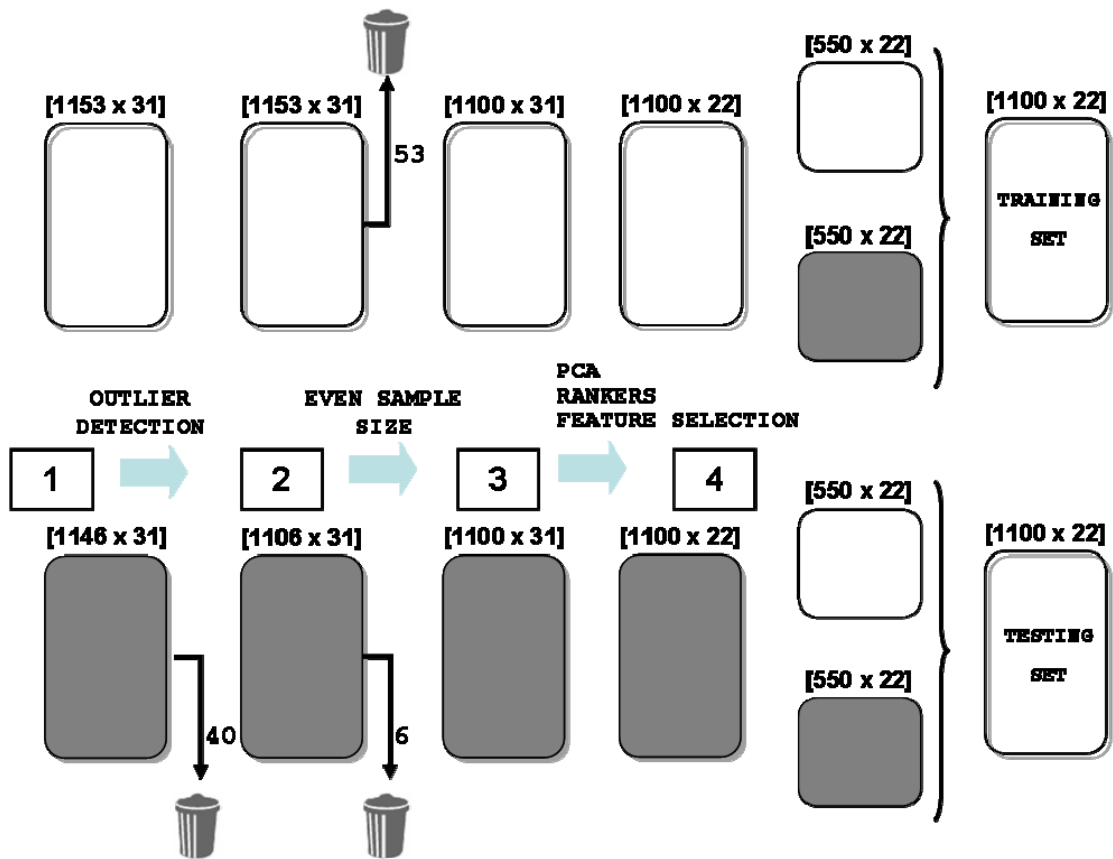


Fig 29. Esquema de la primera etapa de este trabajo. Los pasos que se llevaron a cabo fueron: 1) Detección de outliers y medidas básicas de centralidad y dispersión para los modelos correctos (en blanco) e incorrectos (en gris), 2) Uniformación del número de modelos por clase, 3) Aplicación de los métodos de ranking, selección de variables y extracción de variables, y 4) formación de los conjuntos de entrenamiento (training set) y prueba (testing set).

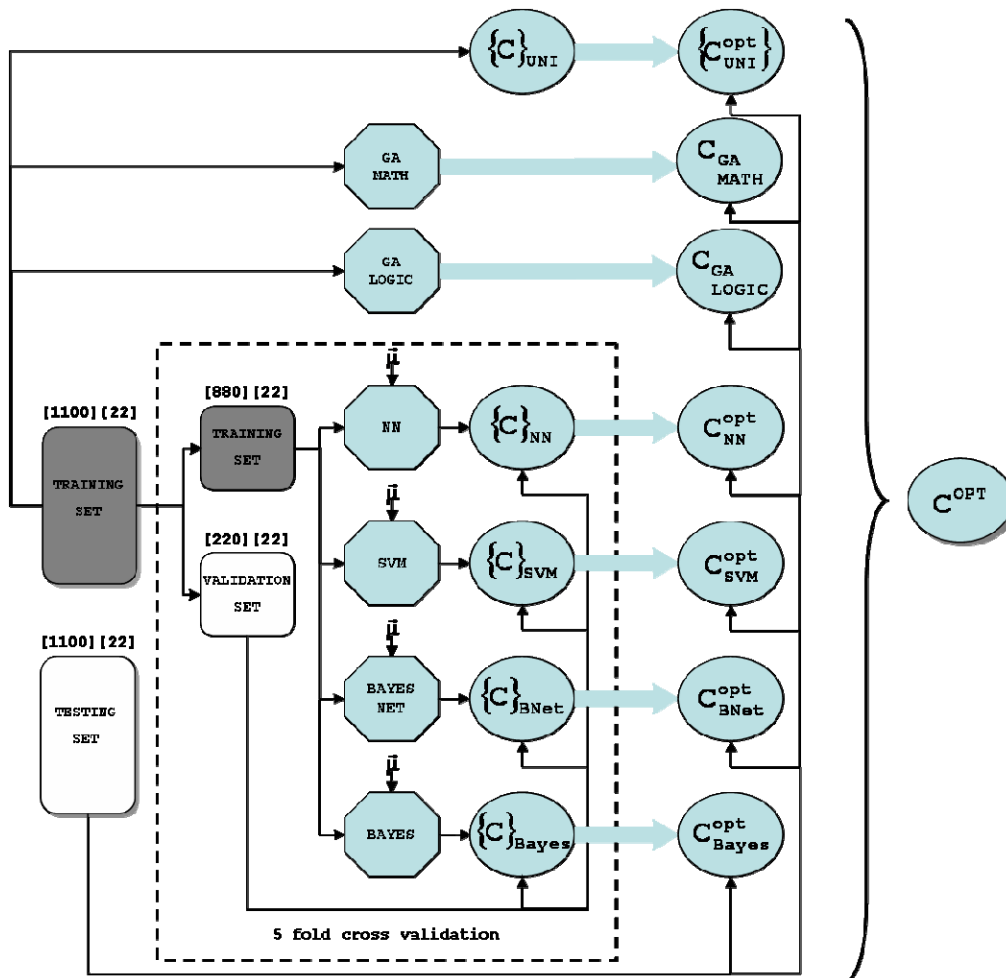


Fig 30. Esquema de la segunda etapa de este trabajo. Los pasos que se llevaron a cabo fueron: 1) entrenamiento y prueba de algoritmos de aprendizaje y clasificadores que no requieren optimización de parámetros, 2) entrenamiento, validación y prueba de los algoritmos de aprendizaje y clasificadores que sí requieren optimización de parámetros, y 3) comparación de los clasificadores obtenidos y determinación de un clasificador óptimo dentro de aquellos generados.

3.2) Medidas básicas y detección de outliers.

Con el fin de entender el dominio de cada variable, el primer paso corresponde al análisis individual de cada una, para los modelos correctos (Tabla 8) e incorrectos (Tabla 9).

Tabla 8. Estadísticos básicos medidos sobre la población de modelos correctos.

VARIABLE	MIN	MAX	RANGO	PROMEDIO	VAR	DV STD	SK	KURT
ALI_seqide	10.81	39.47	28.66	24.81	40.00	6.32	0.36	-0.66
ALI_zsco	-16.53	-1.05	15.48	-8.78	5.76	2.40	0.01	0.68
ALI_pbeval	0.00	100.00	100.00	3.52	151.90	12.32	4.66	24.22
ALI_tgcov	4.83	100.00	95.17	54.70	1071.57	32.73	0.26	-1.58
ML_len	30.00	149.00	119.00	88.83	832.70	28.86	-0.01	-0.74
ML_comp	0.02	0.75	0.72	0.32	0.01	0.11	0.56	0.46
ML_zcomb	-8.98	1.67	10.65	-5.01	2.55	1.60	0.58	0.84
ML_zpair	-7.19	0.95	8.14	-3.72	1.60	1.26	0.23	-0.21
ML_zsurf	-7.02	2.39	9.41	-3.37	1.64	1.28	0.71	1.76
ML_pp	1.27	3.89	2.62	1.86	0.05	0.21	2.47	16.04
ML_coabs	4.00	57.40	53.40	28.36	73.55	8.58	-0.46	-0.20
ML_corel	0.07	0.49	0.42	0.33	0.01	0.07	-0.56	-0.15
ML_rg	9.00	24.07	15.07	13.49	1.99	1.41	-0.08	5.65
TF_len	29.00	149.00	120.00	85.83	762.29	27.61	0.07	-0.57
TF_comp	0.02	0.74	0.71	0.28	0.01	0.10	0.91	1.43
TF_zcomb	-12.18	0.71	12.89	-6.52	3.31	1.82	0.85	1.26
TF_zpair	-11.12	0.23	11.35	-5.26	2.08	1.44	0.22	0.81
TF_zsurf	-7.70	1.69	9.39	-3.96	1.95	1.40	1.00	1.35
TF_pp	1.28	3.85	2.57	1.80	0.05	0.23	3.03	23.15
TF_coabs	4.00	56.10	52.10	27.82	67.07	8.19	-0.59	-0.04
TF_corel	0.07	0.50	0.42	0.33	0.01	0.07	-0.53	-0.15
TF_rg	8.82	24.11	15.28	13.55	2.02	1.42	-0.10	5.88
TW_len	29.00	168.00	139.00	90.28	853.17	29.21	0.06	-0.46
TW_comp	0.02	0.74	0.71	0.29	0.01	0.10	0.90	1.48
TW_zcomb	-11.96	0.92	12.88	-6.79	3.53	1.88	0.96	1.46
TW_zpair	-12.88	0.03	12.91	-5.42	2.10	1.45	0.28	1.22
TW_zsurf	-7.86	1.57	9.43	-4.17	2.13	1.46	0.96	1.41
TW_pp	1.28	3.85	2.57	1.83	0.05	0.23	2.45	17.62
TW_coabs	4.00	57.40	53.40	28.78	69.06	8.31	-0.73	0.14
TW_corel	0.07	0.50	0.42	0.33	0.01	0.07	-0.54	-0.19
TW_rg	8.82	24.11	15.28	13.72	2.13	1.46	-0.20	5.15

VARIABLE: identificador para cada variable medida, tal como se define en Tabla 7. MIN: mínimo valor que toma la variable. MAX: máximo valor que toma la variable. RANGO: distancia entre el mínimo y máximo valor. PROMEDIO: media aritmética de la variable. VAR: varianza de la variable. DV STD: desviación estandar de la variable. SK: skewness de la variable. KURT: kurtosis de la variable.

Tabla 9. Estadísticos básicos medidos sobre la población de modelos incorrectos.

VARIABLE	MIN	MAX	RANGO	PROMEDIO	VAR	DV STD	SK	KURT
ALI_seqide	14.50	86.99	72.49	20.25	37.47	6.12	3.40	23.82
ALI_zsco	-28.07	1.20	29.27	-5.64	6.73	2.59	-2.10	11.34
ALI_pbeval	0.00	99.00	99.00	18.72	768.17	27.72	1.44	0.88
ALI_tgcov	3.12	100.00	96.88	43.91	646.26	25.42	0.55	-0.77
ML_len	30.00	149.00	119.00	86.51	1080.78	32.88	0.23	-1.07
ML_comp	0.01	0.85	0.85	0.24	0.02	0.14	1.04	1.34
ML_zcomb	-11.51	2.77	14.28	-2.43	3.93	1.98	-0.75	1.28
ML_zpair	-9.74	2.56	12.30	-2.02	2.82	1.68	-0.70	1.35
ML_zsurf	-6.18	3.56	9.74	-1.42	1.94	1.39	-0.31	0.57
ML_pp	1.33	5.20	3.87	2.22	0.14	0.38	1.45	5.17
ML_coabs	4.00	63.90	59.90	19.68	71.33	8.45	0.66	0.35
ML_corel	0.03	0.52	0.48	0.24	0.01	0.08	0.31	-0.09
ML_rg	8.68	54.68	46.00	15.27	17.78	4.22	3.07	18.00
TF_len	29.00	147.00	118.00	81.62	921.49	30.36	0.28	-0.95
TF_comp	0.00	0.80	0.80	0.22	0.02	0.14	0.98	0.99
TF_zcomb	-13.07	1.31	14.38	-5.38	4.83	2.20	-0.06	0.32
TF_zpair	-12.70	0.67	13.38	-4.81	3.82	1.95	-0.65	1.32
TF_zsurf	-6.81	2.87	9.68	-2.80	2.29	1.51	0.39	0.15
TF_pp	1.14	5.33	4.19	1.88	0.11	0.32	2.61	16.42
TF_coabs	4.00	63.80	59.80	18.62	63.89	7.99	0.65	0.45
TF_corel	0.03	0.50	0.47	0.24	0.01	0.09	0.30	-0.20
TF_rg	8.51	60.15	51.64	15.65	24.22	4.92	3.16	18.19
TW_len	29.00	183.00	154.00	87.33	1155.36	33.99	0.36	-0.77
TW_comp	0.00	0.80	0.80	0.23	0.02	0.14	0.89	0.79
TW_zcomb	-14.92	1.52	16.44	-5.76	5.56	2.36	-0.16	0.52
TW_zpair	-16.61	0.70	17.31	-5.13	4.69	2.17	-0.94	2.43
TW_zsurf	-6.80	2.62	9.42	-3.02	2.36	1.53	0.46	0.18
TW_pp	1.14	5.33	4.19	1.88	0.10	0.32	2.61	17.28
TW_coabs	4.00	64.30	60.30	19.62	74.15	8.61	0.66	0.28
TW_corel	0.03	0.50	0.47	0.24	0.01	0.08	0.27	-0.22
TW_rg	8.51	60.79	52.28	15.69	23.95	4.89	3.10	17.63

VARIABLE: identificador para cada variable medida, tal como se define en Tabla 7. MIN: mínimo valor que toma la variable. MAX: máximo valor que toma la variable. RANGO: distancia entre el mínimo y máximo valor. PROMEDIO: media aritmética de la variable. VAR: varianza de la variable. DV STD: desviación estandar de la variable. SK: skewness de la variable. KURT: kurtosis de la variable.

Los gráficos de probabilidad normal se presentan en el anexo A de este trabajo. De acuerdo a estos gráficos y a los valores de skewness y kurtosis, no se puede asumir con certeza una distribución normal de las variables.

Los gráficos de caja y bigotes se presentan en el anexo B de este trabajo. De acuerdo a los resultados obtenidos en estos gráficos, y aplicando conocimiento del dominio en que se encuentra cada variable, se decidió eliminar todos los outliers y outliers lejanos para el caso de identidad del alineamiento (ALI_seqide). Esto generó el descarte de 40 modelos incorrectos.

3.3) Rankers

Con el fin de conocer la relevancia de cada variable con respecto a la clase respuesta, se calcularon índices que miden esta relevancia y proporcionan un ranking de las variables de acuerdo a este criterio (Tabla 10).

Tabla 10. Ranking de variables.

RANKING	GRF	RAZON DE ENTROPIA	SU
1	ML_zcomb	ML_zsurf	ML_zsurf
2	ML_zsurf	ML_zcomb	ML_zcomb
3	ALI_zsco	ALI_zsco	ALI_zsco
4	ML_zpair	ML_pp	ML_pp
5	ML_pp	ML_zpair	ML_zpair
6	TW_corel	ML_corel	ML_corel
7	TF_corel	TF_corel	TW_corel
8	ML_corel	TW_corel	TF_corel
9	TF_coabs	TF_coabs	TF_rg
10	TW_coabs	ML_rg	TF_coabs
11	ML_coabs	TW_coabs	ALI_pbeval
12	ALI_seq_ide	TF_rg	TW_coabs
13	TF_zsurf	ML_coabs	TW_rg
14	TW_zsurf	TW_rg	ML_coabs
15	ALI_pbeval	ALI_seq_ide	ML_rg
16	TF_zcomb	TF_comp	ALI_seq_ide
17	TF_rg	ML_comp	TF_comp
18	ML_rg	TW_comp	TW_comp
19	ML_comp	TF_zsurf	ML_comp
20	TW_rg	TW_zsurf	TF_zsurf
21	TF_comp	TF_zcomb	TW_zsurf
22	TW_comp	ALI_tgcov	ALI_tgcov
23	TW_zcomb	TW_zcomb	TF_zcomb
24	ALI_tgcov	ML_len	TW_zcomb
25	TF_pp	TF_len	ML_len
26	TF_zpair	TF_pp	TW_len
27	TW_pp	TW_zpair	TF_len
28	TW_zpair	TW_len	TF_zpair
29	TF_len	TF_zpair	TW_zpair
30	TW_len	TW_pp	TF_pp
31	ML_len	ALI_pbeval	TW_pp

GRF: Grado de Relación Funcional. SU: Incertidumbre Simétrica. Celdas del mismo color identifican variables con las mismas propiedades, independiente de la categoría (modelo, fragmento del molde, o molde completo). RAZON DE ENTROPIA: razón de las entropías intragrupo y entropía total (ver métodos para más detalles).

En el Anexo C se presentan los valores que toma cada variable para cada índice de ranking. En general se observa una coherencia en el ranking entregado por cada índice (Fig 31).

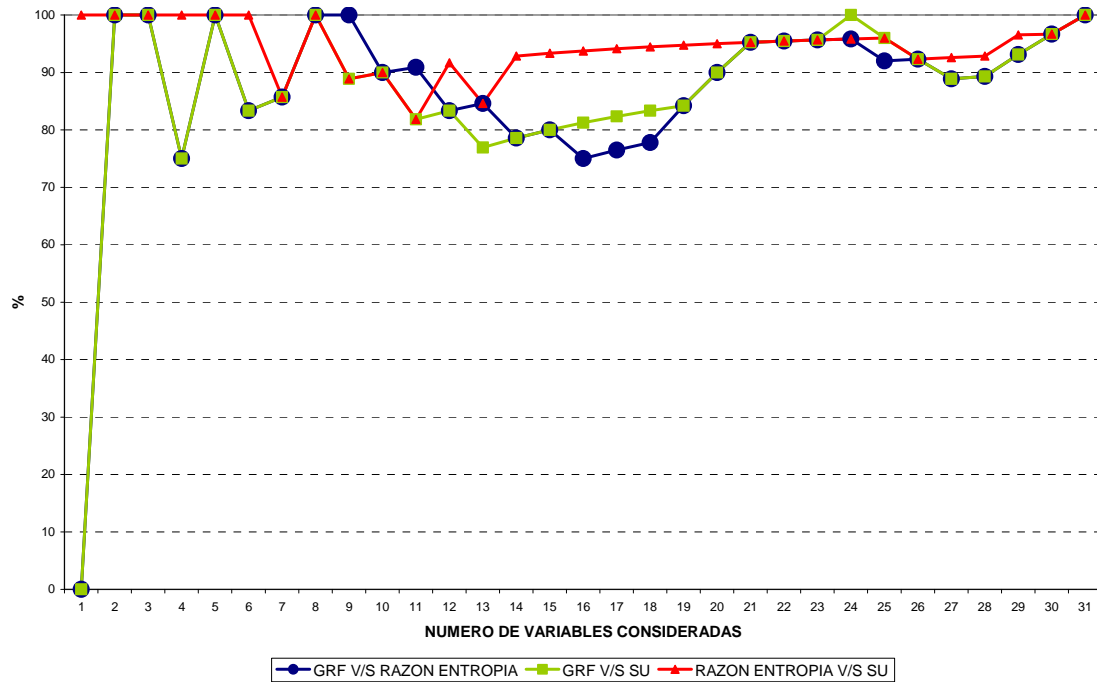


Fig 31. Porcentaje de calce entre los distintos métodos de ranking. Para cada número de variables consideradas, se determina para cada par de métodos cuantas variables están presentes dentro de ese número. Por ejemplo, si se observan las primeras 8 variables rankeadas en cada caso, el 100% de éstas son las mismas entre cada par de métodos.

3.4) Selección de variables

3.4.1) Algoritmo de Yu y Liu.

Este algoritmo se aplicó con el fin de obtener un subconjunto de variables que maximice la relevancia con respecto a la clase respuesta y minimice la redundancia entre las variables. El parámetro umbral δ , que permite definir un conjunto inicial de variables relevantes, se fijó en cero de manera de considerar a todas las variables como inicialmente relevantes para la clase respuesta, y así no descartar ninguna *a priori*.

El algoritmo se ejecutó en un total de cuatro iteraciones (Tabla 11).

Tabla 11. Selección de variables: algoritmo de Yu y Liu.

ITERACION	VARIABLE RELEVANTE (VR)/VARIABLES ELIMINADAS	CONJUNTO FINAL
ITERACION 1	VR: ML_zsurf	ALI_seqide, ML_corel, ML_pp, ML_zsurf
	ML_zcomb, ALI_zsco, ML_zpair, TF_coabs, ALI_pbeval, TW_coabs, ML_coabs, TF_zsurf, TW_zsurf, ALI_tgcov, TF_zcomb, TW_zcomb, ML_len, TW_len, TF_len, TF_zpair, TW_zpair, TF_pp, TW_pp	
ITERACION 2	VR: ML_pp	
ITERACION 3	VR: ML_corel	
	TW_corel, TF_corel, TF_rg, TW_rg, ML_rg, TF_comp, TW_comp, ML_comp	
ITERACION 4	VR: ALI_seqide	

En cada iteración, el algoritmo determina la variable más relevante para la clase respuesta y elimina aquellas variables que son igual o menos relevantes para la clase respuesta, y más redundantes con respecto a aquella más relevante. En esta tabla, para cada iteración (primera columna) se muestra primero la variable más relevante (destacado en negro) y luego las variables eliminadas (segunda columna). Por ejemplo, para la segunda iteración, la variable más relevante es ML_pp y no se eliminan variables en esta iteración. Finalmente, el subconjunto de variables más relevantes es el resultado del algoritmo (tercera columna).

3.4.2) Algoritmo de Koller y Sahami

Este algoritmo se implementó y ejecutó con el fin de determinar un ranking de las variables que considere tanto la redundancia entre ellas como la relevancia con respecto a la clase respuesta. Dependiendo del tamaño del manto de markov utilizado, se obtienen rankings que difieren significativamente (Tabla 12). Sin embargo, para el caso en que no hay manto de markov, se observa una coherencia entre este algoritmo y los métodos de ranking.

Tabla 12. Selección de variables: algoritmo de Koller y Sahami.

RELEVANCIA	K=0	K=1	K=2
1	ML_zsurf	ML_zsurf	ML_zsurf
2	ML_zcomb	TF_corel	ALI_seqide
3	ALI_zsco	ALI_tgcov	TW_len
4	ML_pp	ML_len	ML_comp
5	TF_corel	TF_rg	ML_pp
6	TW_corel	ALI_seqide	TW_coabs
7	ML_corel	ML_pp	TW_pp
8	ML_zpair	ALI_zsco	ALI_tgcov
9	TF_coabs	TF_coabs	TW_zpair
10	TW_coabs	TW_comp	ML_zpair
11	ML_coabs	TF_pp	ALI_zsco
12	TF_rg	TF_zsurf	TW_zsurf
13	ALI_seqide	TF_zcomb	TF_zpair
14	TW_rg	TF_zpair	TW_rg
15	ML_rg	TF_len	TW_corel
16	TW_comp	ML_comp	TF_zsurf
17	ML_comp	ML_zcomb	ML_coabs
18	TF_comp	ML_zpair	TF_zcomb
19	ALI_pbeval	TW_len	TW_zcomb
20	ALI_tgcov	TW_zsurf	TF_comp
21	TF_zsurf	ML_rg	ML_len
22	TW_zsurf	TW_pp	ML_rg
23	TF_zcomb	TW_zpair	TF_pp
24	ML_len	ML_coabs	ML_corel
25	TW_zcomb	TW_coabs	ALI_pbeval
26	TW_len	TW_zcomb	TW_comp
27	TF_len	ALI_pbeval	TF_len
28	TF_zpair	ML_corel	ML_zcomb
29	TW_zpair	TF_comp	TF_coabs
30	TF_pp	TW_corel	TF_corel
31	TW_pp	TW_rg	TF_rg

El valor K determina el número de variables que componen el manto de markov durante la ejecución del algoritmo. El hecho de utilizar un manto de markov igual K=0 es equivalente a utilizar un método de ranking.

3.5) Análisis de componentes principales (ACP).

Con el fin de generar un conjunto de variables nuevas que capturen la información presente en las variables originales y que reduzcan la dimensionalidad, se aplicó la técnica de ACP. La cantidad de información que captura cada componente principal, así como los pesos de cada variable asociados a cada componente, se muestran en el Anexo D. En general, existe una alta correlación en la participación de las variables que corresponden a una misma propiedad con respecto a las nuevas componentes principales (Fig 32 y Fig 33). Las primeras nueve componentes principales capturan un 90% de la información de las variables originales (Fig 34).

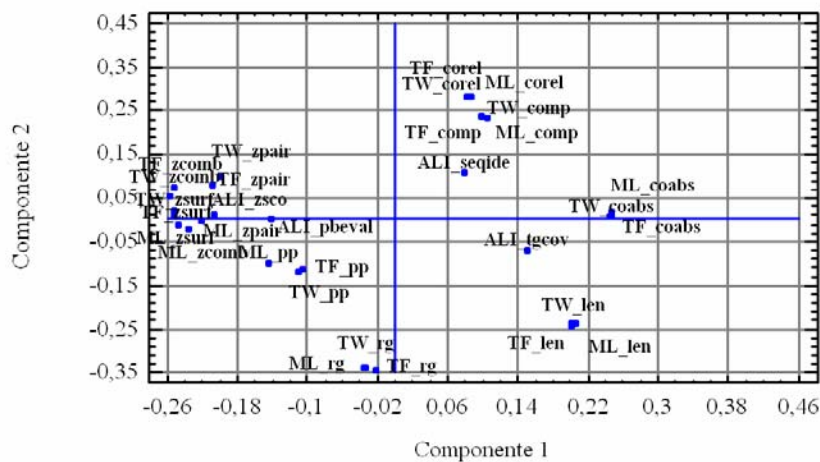


Fig 32. Variables en la primera y segunda componente principal. Participación de cada variable en las primeras dos componentes principales, que capturan el 59.1% de la información. Como se observa, las variables asociadas a una misma propiedad tienden a agruparse, mostrando el alto nivel de correlación entre ellas.

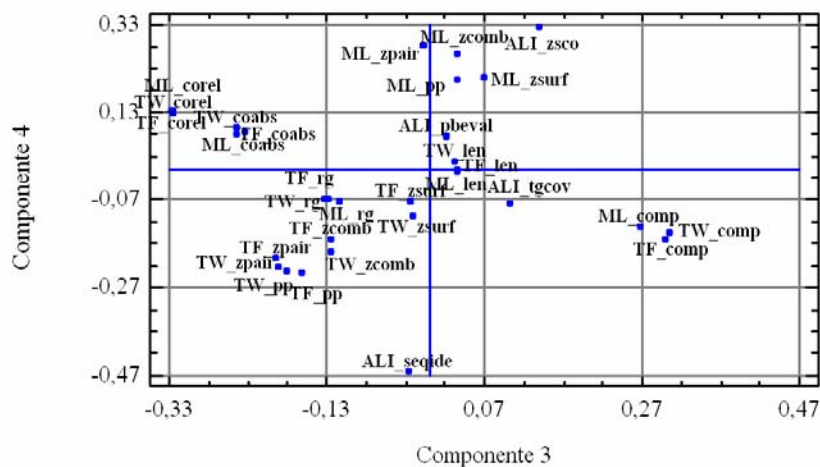


Fig 33. Variables en la tercera y cuarta componente principal. Participación de cada variable en la tercera y cuarta componente principal, que juntas capturan un 15.4% de la información.

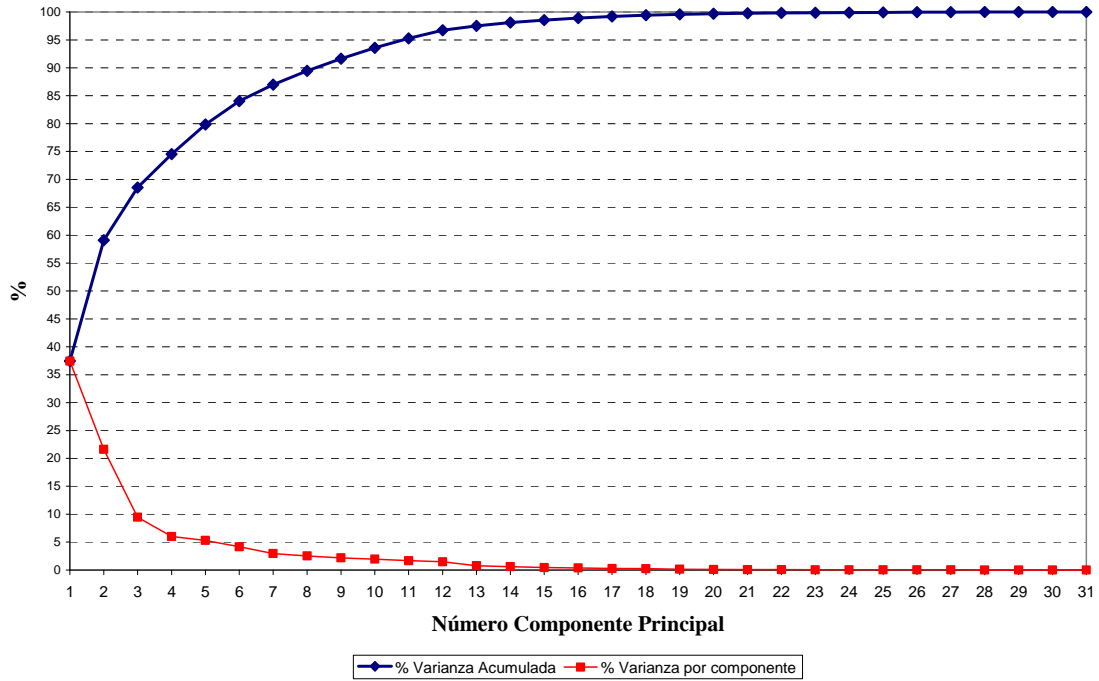


Fig 34. Varianza capturada por componente principal, y su distribución acumulada.

Dados los resultados obtenidos con los métodos de ranking, selección de variables y ACP, en los cuales se evidencia una alta redundancia entre las propiedades medidas sobre el molde completo para generar el modelo proteico (TW) y aquellas medidas sobre la fracción efectiva del molde utilizado para generar el modelo (TF), se decidió eliminar del conjunto inicial de variables a todas aquellas pertenecientes a la categoría TW. Así, se tiene un total de 22 variables que serán utilizadas para entrenar los algoritmos de aprendizaje que generan los clasificadores a evaluar en este trabajo.

3.6) Clasificadores.

A continuación se presentan los resultados de los diferentes clasificadores generados en este trabajo. Primero, se muestran los resultados de las variables individuales como clasificadores, lo que permite generar una idea de cual es el rendimiento que un clasificador multivariable debe superar de manera de ser considerado como útil. Luego, se muestran los resultados de los clasificadores que no requieren optimización de parámetros. Finalmente, se muestran los resultados de aquellos clasificadores que fueron generados por medio de la optimización de uno o más parámetros.

3.6.1) Variables como clasificadores.

Cada variable medida sobre los modelos proteicos fue utilizada como un clasificador. Para cada una de ellas, se midió la precisión y área bajo la curva ROC (AUC), tanto sobre el conjunto de entrenamiento como el de prueba (Tabla 13).

Tabla 13. Precisión y AUC de cada variable sobre los conjuntos de entrenamiento y prueba.

VARIABLE	OT	ACC TR	ACC TE	AUC TR	AUC TE
ML_ZCOMB	-3.53	82.5	80.2	0.876	0.846
ALI_ZSCO	-6.71	79.5	78.9	0.86	0.843
ML_ZSURF	-2.67	85.6	79.5	0.89	0.84
ML_PP	1.99	78.3	76.4	0.841	0.825
ML_ZPAIR	-2.64	74.8	74.2	0.814	0.808
TF_COREL	0.33	75.3	73.8	0.797	0.786
TW_COREL	0.29	75.1	71.8	0.802	0.786
ML_COREL	0.31	75.4	74.4	0.796	0.784
ALI_SEQIDE	19.39	70	70.7	0.752	0.764
TF_COABS	23.6	78.2	73.6	0.811	0.758
TW_COABS	25.6	78.3	74	0.804	0.744
ALI_PBEVAL	0.05	72.8	68.8	0.78	0.735
ML_COABS	24	77.4	72.5	0.791	0.735
TF_ZSURF	-3.6	70.9	69.2	0.75	0.72
TW_ZSURF	-3.72	71	67.5	0.747	0.706
ML_COMP	0.2	66.4	65.8	0.685	0.68
TF_COMP	0.16	66.8	65	0.667	0.671
TF_ZCOMB	-5.68	66.7	64.7	0.69	0.664
TW_COMP	0.17	68	65.7	0.658	0.661
ML_RG	14.89	68.2	67.8	0.634	0.653
TF_RG	15.12	68.8	67.6	0.633	0.644
TW_ZCOMB	-6.27	66	62.9	0.673	0.642
TW_RG	15.53	67.5	67.2	0.612	0.629
TF_ZPAIR	-4.47	61.3	60	0.607	0.595
TF_PP	1.99	57.6	56	0.553	0.572
TW_ZPAIR	-4.72	61.1	58.7	0.585	0.572
ALI_TGCOV	88.48	65.3	60.6	0.623	0.556
TW_PP	2.08	55.3	55.1	0.518	0.549
TF_LEN	59	59.9	53.5	0.564	0.531
TW_LEN	71	60.3	54.3	0.557	0.516
ML_LEN	62	59.6	55	0.546	0.51

Las variables se muestran ordenadas de mayor a menor AUC sobre el conjunto de prueba. OT: umbral óptimo (score) obtenido sobre el conjunto de training; ACC TR: precisión al aplicar el OT sobre el conjunto de entrenamiento. ACC TE: precisión al aplicar el OT sobre el conjunto de prueba. AUC TR: área bajo la curva ROC sobre el conjunto de entrenamiento. AUC TE: área bajo la curva ROC sobre el conjunto de prueba.

Tanto en términos de precisión como de AUC, las propiedades correspondientes al Z-score de energía combinado del modelo (ML_zcomb), al Z-score del alineamiento secuencia-estructura (ALI_zsco) y al Z-score de energía de superficie accesible del modelo (ML_zsurf) son aquellas de mejor rendimiento.

3.6.2) Clasificador basado en distancia a centroides

Para aplicar el clasificador basado en distancia a centroides es necesario estimar los mismos, lo que corresponde a calcular el vector de promedios de las variables en el conjunto de entrenamiento, tanto para el grupo correcto como incorrecto (Tabla 14).

Tabla 14. Centroides estimados para las clases correcto e incorrecto.

VARIABLE	PROM. INCORRECTO	PROM. CORRECTO
ALI_SEQIDE	19.51	24.57
ALI_ZSCO	-5.41	-8.87
ALI_PBEVAL	19.41	2.47
ALI_TGCOV	43.00	56.92
ML_LEN	86.42	90.33
ML_COMP	0.24	0.32
ML_ZCOMB	-2.31	-5.03
ML_ZPAIR	-1.95	-3.70
ML_ZSURF	-1.33	-3.43
ML_PP	2.24	1.85
ML_COABS	19.66	29.10
ML_COREL	0.24	0.33
ML_RG	15.19	13.53
TF_LEN	81.62	87.15
TF_COMP	0.22	0.28
TF_ZCOMB	-5.39	-6.59
TF_ZPAIR	-4.83	-5.29
TF_ZSURF	-2.81	-4.04
TF_PP	1.87	1.80
TF_COABS	18.59	28.48
TF_COREL	0.24	0.34
TF_RG	15.55	13.60

Los resultados de aplicar el clasificador basado en distancia a centroides sobre el conjunto de entrenamiento entregan una precisión del 59.0% y un área bajo la curva ROC de 0.702. Al aplicarlo sobre el conjunto de prueba, se obtiene una precisión del 56.3% y un área bajo la curva ROC de 0.645. Así, medir la distancia de cada modelo del conjunto de entrenamiento y de prueba a los centroides resulta en un método de clasificación que tiene un rendimiento bajo tanto en términos de precisión como de AUC.

3.6.3) Algoritmos genéticos.

3.6.3.1) GA math.

El algoritmo genético GA Math se ejecutó en 100,000 iteraciones independientes para valores de profundidad máxima del árbol que representa cada función de 3, 4, 5 y 6; los valores de los distintos parámetros se mantuvieron fijos (Tabla 5). Aquella fórmula que exhibe un mejor rendimiento tanto en términos de precisión como de AUC sobre el conjunto de prueba corresponde a la generada con una profundidad máxima de 6 (Tabla 15).

Tabla 15. Precisión y AUC sobre los conjuntos de entrenamiento y prueba para los clasificadores GA Math.

MAX PROF	TP*TN TR	TP*TN TE	ACC TR	ACC TE	OT	AUC TR	AUC TE
3	0.813	0.784	90.2	88.5	7.351	0.954	0.933
4	0.838	0.805	91.5	89.7	8.970	0.958	0.946
5	0.842	0.810	91.8	90.0	12.06	0.968	0.954
6	0.832	0.811	91.3	90.1	10.09	0.964	0.954

TP*TN TR: producto de la tasa de verdaderos positivos y la tasa de verdaderos negativos sobre el conjunto de entrenamiento. TP*TN TE: producto de la tasa de verdaderos positivos y la tasa de verdaderos negativos sobre el conjunto de entrenamiento. OT: umbral óptimo (score) obtenido sobre el conjunto de training. ACC TR: precisión al aplicar el OT sobre el conjunto de entrenamiento. ACC TE: precisión al aplicar el OT sobre el conjunto de prueba. AUC TR: área bajo la curva ROC sobre el conjunto de entrenamiento. AUC TE: área bajo la curva ROC sobre el conjunto de prueba.

Las fórmulas obtenidas para cada caso se presentan a continuación.

Máxima Profundidad = 3

$$ML_RG + ML_ZCOMB + ML_COREL \cdot ALI_ZSCO$$

Máxima Profundidad = 4

$$ML_PP \cdot (ALI_ZSCO + ML_RG) - ML_ZCOMB \cdot ML_ZPAIR \cdot TF_COREL$$

Máxima Profundidad = 5

$$ML_COREL^2 \cdot ML_RG \cdot ML_ZCOMB + ML_COMP \cdot TF_RG + (ALI_ZSCO + ML_RG) \cdot |ML_PP|$$

Máxima Profundidad = 6

$$(ALI_ZSCO + ML_RG + TF_COMP) \cdot (ML_PP + ML_ZCOMB \cdot TF_COREL^2)$$

Con el fin de entender mejor la importancia de cada variable, se analizó el número de fórmulas (de las 100,000 soluciones arrojadas por este GA) en las que cada variable aparece, independiente del número de veces que ocurra por fórmula (Fig 35).

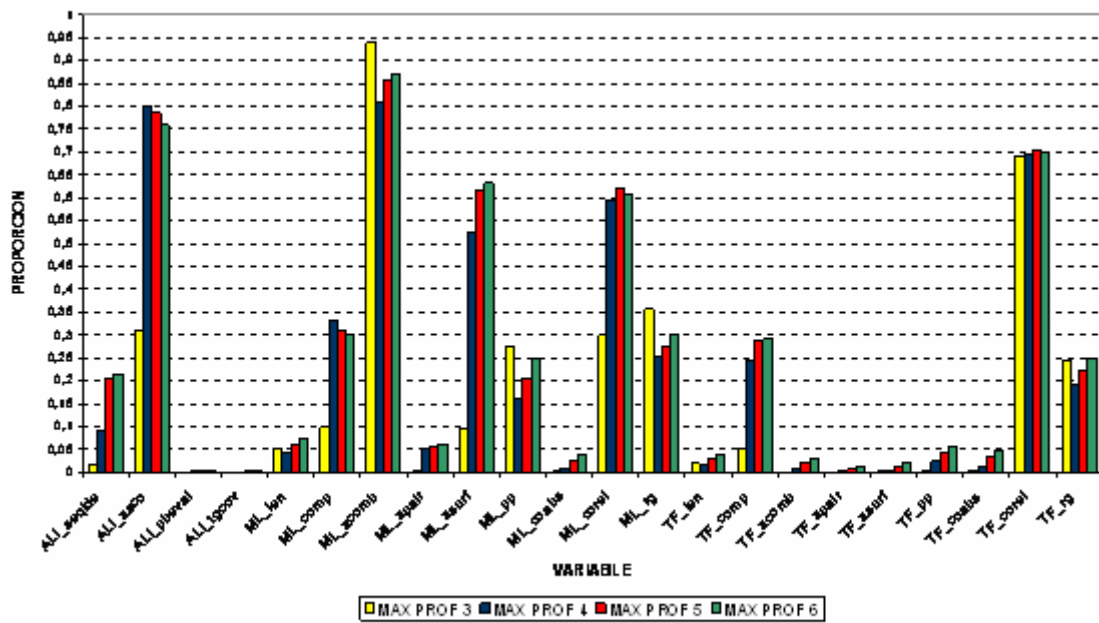


Fig 35. Frecuencia de ocurrencia de cada variable en las fórmulas de GA Math. Proporción de veces que aparece cada variable con respecto a las 100,000 fórmulas generadas por GA Math para cada valor de profundidad. Esta proporción sólo considera el número de fórmulas en las que aparece cada variable independientemente del número de veces que aparece cada variable en cada fórmula.

3.6.3.2) GA logic

El algoritmo genético GA Logic se ejecutó en 100,000 iteraciones independientes para un número de genes desde 2 hasta 10; los valores de los distintos parámetros se mantuvieron fijos (Tabla 4).

Aquella fórmula lógica que exhibe un mejor rendimiento en términos de precisión corresponde a aquella generada con un número de genes igual a 9. En términos de AUC, la fórmula con mejor rendimiento corresponde a aquella con un número de genes igual a 10 (Tabla 16).

Tabla 16. Precisión y AUC sobre los conjuntos de entrenamiento y prueba para cada clasificador obtenido con GA Logic.

NUM GEN	ACC TR	ACC TE	AUC TR	AUC TE
2	87.3	84.6	0.877	0.850
3	90.2	86.3	0.922	0.891
4	89.8	87.2	0.925	0.906
5	91.0	87.4	0.932	0.888
6	90.7	88.5	0.932	0.904
7	89.6	88.2	0.922	0.907
8	91.1	88.5	0.939	0.907
9	90.6	88.6	0.928	0.908
10	90.3	88.4	0.928	0.917

NUM GEN: número de genes que componen a cada fórmula lógica óptima. ACC TR: precisión al aplicar el clasificador sobre el conjunto de entrenamiento. ACC TE: precisión al aplicar el clasificador sobre el conjunto de prueba. AUC TR: área bajo la curva ROC sobre el conjunto de entrenamiento. AUC TE: área bajo la curva ROC sobre el conjunto de prueba.

Los árboles de decisión obtenidos se pueden representar por las siguientes fórmulas lógicas, de acuerdo al número de genes.

Número de Genes = 2

$$(\mathbf{ML_ZCOMB} \leq -3.23 \wedge \mathbf{TF_RG} \leq 16.31)$$

Número de Genes = 3

$$((\mathbf{ALI_SEQIDE} > 31.98) \vee (\mathbf{ML_ZSURF} \leq -2.59 \wedge \mathbf{TF_RG} \leq 16.15))$$

Número de Genes = 4

$$\neg((\mathbf{ALI_ZSCO} > -7.82 \wedge \mathbf{ML_LEN} > 35.65 \wedge \mathbf{ML_ZSURF} > -2.85) \vee (\mathbf{TF_RG} > 15.66))$$

Número de Genes = 5

$$\neg \left(\left((\mathbf{TF_RG} > 15.74) \vee (\mathbf{ML_ZSURF} > -2.17 \wedge \mathbf{TF_LEN} > 42.36) \right) \vee (\mathbf{ALI_ZSCO} > -5.07 \wedge \mathbf{ALI_TGCOV} > 28.61) \right)$$

Número de Genes = 6

$$\neg \left(\left((\mathbf{ALI_TGCOV} > 12.66 \wedge \mathbf{ALI_ZSCO} > -5.35) \vee (\mathbf{TF_RG} > 16.28) \right) \vee (\mathbf{ALI_PBEVAL} \leq 108.84 \wedge \mathbf{ML_LEN} > 47.47 \wedge \mathbf{ML_ZSURF} > -2.11) \right)$$

Número de Genes = 7

$$\neg \left(\left((\mathbf{ALI_ZSCO} > -10.49 \wedge \mathbf{TF_COMP} > 1.48) \vee (\mathbf{ML_LEN} > 46.42 \wedge \mathbf{ML_ZCOMB} > -2.99) \right) \vee (\mathbf{TF_RG} > 16.62) \vee (\mathbf{ALI_ZSCO} > -5.28 \wedge \mathbf{ALI_TGCOV} > 15.59) \right)$$

Número de Genes = 8

$$\left(\begin{array}{l} (ALI_TGCOV > 11.99 \wedge TF_COREL \leq 0.51 \wedge ALI_ZSCO > -5.77) \vee \\ (ML_PP \leq 2.32 \wedge ML_ZSURF > 45.16) \vee \\ (ML_ZSURF > -2.12 \wedge TF_LEN > 44.92) \vee (TF_RG > 15.93) \end{array} \right)$$

Número de Genes = 9

$$\left(\begin{array}{l} (TF_LEN > 45.94 \wedge ML_ZCOMB \leq 4.88 \wedge ML_ZCOMB > -3.13) \vee (TF_RG > 16.31) \\ \vee (ALI_TGCOV > 14.08 \wedge ALI_ZSCO > -5.77) \vee (TF_ZSURF > 1.31) \vee \\ (ALI_SEQIDE \leq 35.92 \wedge ML_ZCOMB > -0.52) \end{array} \right)$$

Número de Genes = 10

$$\left(\begin{array}{l} (TF_LEN > 155) \vee (ALI_SEQIDE > 32.27) \vee \\ (TF_RG \leq 16.57 \wedge TF_RG \leq 17.45 \wedge ALI_ZSCO \leq -4.96 \wedge ML_ZSURF \leq -2.46) \vee \\ (ALI_TGCOV \leq 12.93 \wedge ML_LEN \leq 43.82) \vee (ALI_PBEVAL > 106.03) \vee (ALI_TGCOV \leq 0.47) \end{array} \right)$$

Con el fin de entender mejor la importancia de cada variable, se analizó el número de fórmulas (de las 100,000 soluciones arrojadas por este GA) en las que cada variable aparece, independiente del número de veces que ocurra por fórmula (Fig 36).

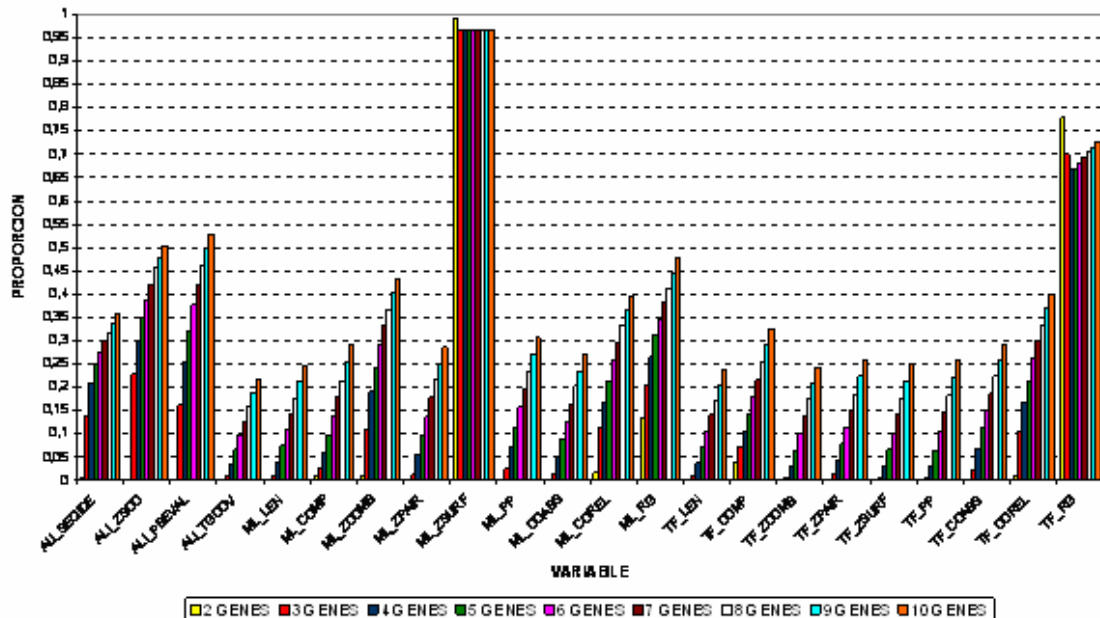


Fig 36. Frecuencia de ocurrencia de cada variable en las reglas lógicas de GA Logic. Proporción de veces que aparece cada variable con respecto a las 100,000 reglas lógicas (o árboles de decisión) generadas por GA Logic, para cada número de genes. Esta proporción sólo considera el número de reglas lógicas en las que aparece cada variable independientemente del número de veces que aparece cada variable en cada regla lógica.

Los algoritmos de aprendizaje que se presentan continuación tienen uno o más parámetros que deben ser optimizados con el fin de generar un clasificador que sea óptimo dentro de los posibles para este algoritmo de aprendizaje. Para lograr esto, se procedió a separar aleatoriamente el conjunto de entrenamiento, que contiene 1,100 modelos, en dos conjuntos:

- un conjunto de entrenamiento propiamente tal, con 880 modelos,
- otro de validación, con 220 modelos.

Estos dos conjuntos se generaron y utilizaron con la modalidad de validación cruzada de orden 5, con el fin de evitar sesgos en la optimización de parámetros proveniente de la separación del conjunto de entrenamiento en aquellos ya mencionados.

3.6.4) Naive Bayes

El clasificador naive Bayes se probó en dos modalidades:

- asumiendo que las variables siguen una distribución normal,
- discretizando las variables, de modo de no asumir una distribución sobre ellas.

Esto se llevó a cabo en modo de validación cruzada de orden 5, obteniéndose una precisión promedio de 89.09% para el primer caso y de 89.91% para el segundo, por lo que se decidió utilizar la modalidad discreta.

Los resultados de entrenar sobre el conjunto de entrenamiento y evaluar tanto sobre este conjunto como en el de prueba el clasificador naive Bayes usando discretización de las variables, corresponden a una precisión sobre el conjunto de entrenamiento de 90.55% y un área bajo la curva ROC de 0.939. Los resultados sobre el conjunto de prueba corresponden a una precisión de 85.18% y un área bajo la curva ROC de 0.900. Así, este clasificador tiene un rendimiento medio.

3.6.5) Perceptrón multicapa (MLP)

Con el fin de lograr una optimización de los pesos de la red que no se sobreajuste al conjunto de entrenamiento, en cada ciclo de aprendizaje del MLP, éste fue entrenado con 660 modelos correspondientes al conjunto de entrenamiento, y el rendimiento de la red fue evaluado sobre los restantes 220 modelos del conjunto de entrenamiento.

Del espacio de valores posibles de parámetros, se muestrearon distintos valores (Tabla 6) con el fin de encontrar aquella combinación de parámetros que genera el mejor clasificador en términos de su precisión (Tabla 17).

Además, se permite un máximo de 5,000 ciclos de aprendizaje y se utiliza un máximo de 20 ciclos en los cuales, si el error aumenta sostenidamente sobre el conjunto de 220 modelos con los cuales el rendimiento de la red es evaluado, se detiene el proceso de aprendizaje.

Tabla 17. Valores optimizados para una y dos capas ocultas en la arquitectura del MLP.

Parámetro	1 Capa Oculta	2 Capas Ocultas
Número de Neuronas primera capa oculta	12	12
Número de Neuronas segunda capa oculta	N.A.	12
Tasa de aprendizaje	0.1	0.4
Tasa de Momentum	0.0-0.1-0.2	0.1

Las arquitecturas de los casos óptimos corresponden a una capa oculta con 12 neuronas, y dos capas ocultas, con 12 neuronas en cada capa; se elige a aquella con la arquitectura más simple como el clasificador con parámetros óptimos para este algoritmo de aprendizaje (Fig 37).

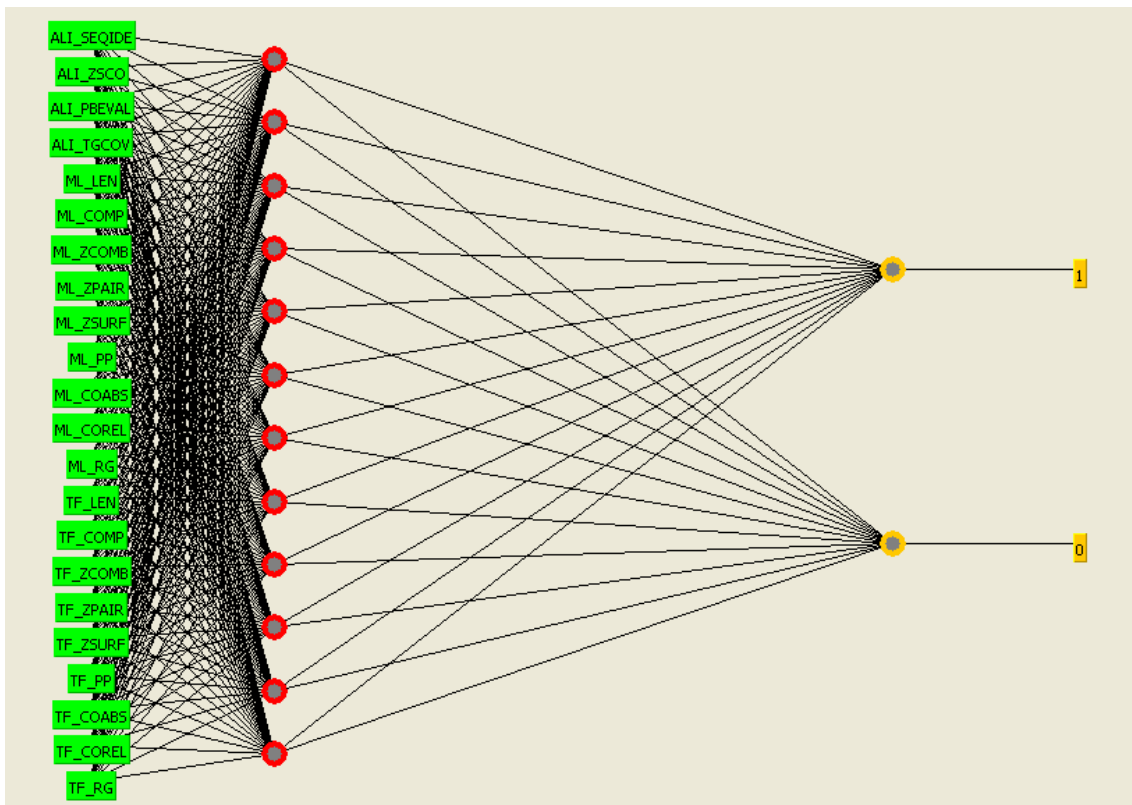


Fig 37. Arquitectura del MLP óptimo con una capa oculta.

Los resultados de entrenar con el conjunto de entrenamiento el clasificador obtenido con el algoritmo de aprendizaje MLP con una capa oculta compuesta por 12 neuronas, corresponden a una precisión sobre el conjunto de entrenamiento de 92.7% y un área bajo la curva ROC de 0.978. Los resultados sobre el conjunto de prueba corresponden a una precisión de 87.73% y un área bajo la curva ROC de 0.956. Así, el rendimiento del clasificador generado tanto en términos de precisión como de AUC es satisfactorio.

Los valores ajustados de los pesos para el clasificador MLP se muestran en el Anexo G de este trabajo.

3.6.6) Redes Bayesianas

Este algoritmo de aprendizaje puede ser utilizado con diferentes algoritmos de búsqueda sobre el espacio de posibles arquitecturas, por lo que es necesario realizar una optimización de estos. Los casos considerados fueron los siguientes:

- Algoritmo K2, con un máximo de 2 padres por nodo, partiendo con una estructura inicial tipo naive Bayes.
- Algoritmo Hill Climbing, con un máximo de 2 padres por nodo, partiendo con una estructura inicial tipo naive Bayes y permitiendo inversiones en el sentido de los arcos.
- Algoritmo TAN, y
- Algoritmo de Annealing Simulado, con $\tau_0 = 10.0$, $\delta = 0.999$ y con un total de iteraciones de 10,000.

Los resultados de la optimización del algoritmo de búsqueda en cuanto a los distintos algoritmos de búsqueda utilizados indican que aquél de mejor rendimiento es TAN (Tabla 18).

Tabla 18. Optimización de algoritmos de búsqueda para Redes Bayesianas.

Algoritmo de búsqueda	ACC Promedio (%)
K2	90.36
Hill Climbing	90.55
TAN	91.00
Annealing Simulado	89.82

La precisión promedio de los distintos algoritmos de búsqueda para redes bayesianas sobre los conjuntos de validación indican que el algoritmo TAN (Tree Augmented Naive Bayes) es aquél de mejor rendimiento.

Los resultados de entrenar con el conjunto de entrenamiento el clasificador generado con el algoritmo de aprendizaje de redes bayesianas con una búsqueda TAN sobre el espacio de posibles arquitecturas, corresponden a una precisión sobre el conjunto de entrenamiento de 93.00%, y a un área bajo la curva ROC de 0.979. Los resultados sobre el conjunto de prueba corresponden a una precisión de 89.36% y a un área bajo la curva de 0.958. Así, el rendimiento del clasificador generado tanto en términos de precisión como de AUC es satisfactorio.

3.6.7) Máquinas de vectores de soporte

Las funciones núcleo consideradas para SVM tienen parámetros que afectan de manera significativa el rendimiento de los clasificadores generados y por lo tanto éstos deben ser optimizados. Recordemos que las funciones núcleo son las siguientes:

- Lineal $\mathbf{x}^t \mathbf{y}$
- Polinómica $(\gamma \cdot \langle \mathbf{x}, \mathbf{y} \rangle + \mathbf{r})^d$

- Gaussiana $\exp(-\gamma\|\mathbf{x} - \mathbf{y}\|^2)$
- Sigmoidal $\tanh(\gamma\langle \mathbf{x}, \mathbf{y} \rangle + \mathbf{r})$

Además, la máquina de vectores de soporte utilizada es aquella de margen blando, por lo que el parámetro que regula la holgura en el error de clasificación (costo C) es otro parámetro a optimizar. Del espacio de valores posibles de parámetros (Tabla 2), se muestrearon distintos valores con el fin de encontrar aquella combinación de parámetros que genera el mejor clasificador en términos de su precisión (Tablas 2 y 3; Anexo E). Cabe mencionar que los valores muestreados fueron determinados previa exploración exhaustiva del espacio de valores posibles de los parámetros.

Tabla 19. Valores optimizados para cada tipo de kernel en SVM.

KERNEL	d	γ	r	C	ACC prom
Lineal	N.A.	N.A.	N.A.	[37.9-38.7]U{88.1;88.3;89.0;89.2;89.6;89.8;90.0;90.2;90.4}	93.09
Polinomial	3	0.1	10	1.5	94.18
	3	0	5	[14.1-15.2]U[19.8-20.0]	94.18
	4	0	10	0.5	94.18
	4	0	5	1.4	94.18
Gaussiana	N.A.	0.9	N.A.	3.6	94.82
Sigmoidal	N.A.	2.6	-10	[0.6-0.9]	88.55

Gráficos de los espacios de parámetros para cada kernel se muestran en el Anexo D de este trabajo. El parámetro **d** representa el exponente en la función kernel polinómica. γ es un factor de ponderación de la medida aplicada sobre dos puntos x e y . **r** corresponde a un desplazamiento en el espacio. **C** es el parámetro que controla las variables de holgura.

Así, se entrenó al algoritmo de aprendizaje con un kernel Gaussiano, con parámetros $\gamma = 0.9$ y $C = 3.6$, utilizando el conjunto de entrenamiento. El clasificador que se obtuvo es:

$$\mathbf{f}(\mathbf{x}) = \text{signo}(\mathbf{h}(\mathbf{x}))$$

con

$$\mathbf{h}(\mathbf{x}) = \sum_{i=1}^n \mathbf{y}_i \alpha_i^* \mathbf{K}(\mathbf{x}_i, \mathbf{x}) + \mathbf{b}^*$$

Donde \mathbf{b}^* es tal que $\mathbf{y}_i \cdot \mathbf{h}(\mathbf{x}_i) = 1$ para todo i que cumple $0 \leq \alpha_i^* \leq C$.

Los valores de cada $\mathbf{y}_i \alpha_i^*$ y cada vector de soporte \mathbf{x}_i (recordemos que aquellos \mathbf{x}_i que no son vectores de soporte tienen un valor $\alpha_i^* = 0$), además del valor para \mathbf{b}^* se encuentran en el Anexo F.

Para generar las curvas ROC se utilizó el valor de decisión de cada modelo \mathbf{x}_i (el modelo es representado como un vector de tamaño 22), que corresponde a $\mathbf{h}(\mathbf{x}_i)$.

Los resultados de entrenar con el conjunto de entrenamiento el clasificador obtenido con el algoritmo de aprendizaje SVM con kernel gaussiano y parámetros $\gamma = 0.9$ y $C = 3.6$, corresponden a una precisión sobre el conjunto de entrenamiento de 99.6% y un área bajo la curva ROC de 0.999. Los resultados sobre el conjunto de prueba corresponden a

una precisión de 93.0% y un área bajo la curva ROC de 0.977. Así, el rendimiento del clasificador generado tanto en términos de precisión como de AUC es satisfactorio.

4) Comparación de clasificadores

Para comparar el rendimiento de los clasificadores óptimos generados con cada técnica evaluada, se aplicó el test de McNemar para comparar las diferencias en las tasas de error y el test de Delong para comparar las diferencias de AUC entre cada par de clasificadores. Si bien la precisión y el AUC son las medidas elegidas para evaluar el rendimiento de los clasificadores, es importante destacar que el AUC es una medida más robusta que la precisión, pues es independiente de un umbral (o threshold) específico que separa las clases en correcto e incorrecto, resumiendo en un solo índice el rendimiento global sobre todo el rango posible de umbrales.

4.1) Test de McNemar.

El resultado de medir si la diferencia entre sus tasas de error es estadísticamente significativa para los 5 mejores atributos en términos de su precisión cuando fueron utilizados como clasificadores (Tabla 13), indica que aquellos atributos basados en propiedades estadísticas son los que exhiben un mejor rendimiento (Tabla 20).

Tabla 20. Resultados del Test de McNemar para la comparación de clasificadores univariados.

	ML_ZCOMB	ML_ZSURF	ALI_ZSCO	ML_PP	ML_COREL
ML_ZCOMB	N.A.	7	14	42	64
ML_ZSURF	0.582	N.A.	7	35	57
ALI_ZSCO	0.338	0.650	N.A.	28	50
ML_PP	0.016	0.054	0.138	N.A.	22
ML_COREL	0.001	0.003	0.011	0.301	N.A.

Este test evalúa la significancia estadística de la diferencia de las tasas de error entre cada uno de los clasificadores. En la tabla, el resultado que se muestra es aquél para las mejores cinco variables como clasificadores. En el triángulo superior derecho de la tabla se muestra el valor absoluto de la diferencia de las tasas de error, para cada par de clasificadores. En el triángulo inferior de la tabla se muestra el p-valor asociado a cada test, para cada par de clasificadores. Un p-valor menor o igual a 0.05 indica que se rechaza la hipótesis nula de igualdad entre ambas tasas de error. Ver Tabla 7 para una descripción del identificador de cada variable.

Para el caso de los clasificadores multivariados óptimos previamente generados, en cuanto a su precisión (Tabla 21), el resultado de medir si la diferencia entre sus tasas de error es estadísticamente significativa indica que SVM es aquél de mejor rendimiento entre los clasificadores generados (Tabla 22).

Tabla 21. Precisiones obtenidas por cada clasificador óptimo multivariable.

CLASIFICADOR	ACC
SVM	93.00
GA_MATH_6	90.09
GA_MATH_5	90.00
GA_MATH_4	89.73
BAYES_NET	89.36
GA_MATH_3	88.55
GA_LOGIC_9	88.55
GA_LOGIC_6	88.45
GA_LOGIC_8	88.45
GA_LOGIC_10	88.36
GA_LOGIC_7	88.18
NN_1LAYER	87.73
NN_2LAYERS	87.55
GA_LOGIC_5	87.36
GA_LOGIC_4	87.18
GA_LOGIC_3	86.27
nBAYES	85.18
GA_LOGIC_2	84.64
CENTROIDS	56.27

ACC: precisión. SVM: Support Vector Machines. GA: Algoritmo genético. NN: Perceptrón Multicapa. BAYES_NET: Red Bayesiana.

Tabla 22. Resultados del Test de McNemar para la comparación de clasificadores multivariables.

	SVM	GA_MATH_6	GA_MATH_5	GA_MATH_4	BAYES_NET	GA_MATH_3	GA_LOGIC_9	GA_LOGIC_6	GA_LOGIC_8	GA_LOGIC_10	GA_LOGIC_7	NN_1LAYER	GA_LOGIC_5	GA_LOGIC_4	GA_LOGIC_3	nBAYES	GA_LOGIC_2	CENTROIDS
SVM	N.A.	32	33	36	40	49	49	50	50	51	53	58	62	64	74	86	92	404
GA_MATH_6	0,001	N.A.	1	4	8	17	17	18	18	19	21	26	30	32	42	54	60	372
GA_MATH_5	0,001	1,000	N.A.	3	7	16	16	17	17	18	20	25	29	31	41	53	59	371
GA_MATH_4	0,000	0,643	0,779	N.A.	4	13	13	14	14	15	17	22	26	28	38	50	56	368
BAYES_NET	0,000	0,480	0,525	0,769	N.A.	9	9	10	10	11	13	18	22	24	34	46	52	364
GA_MATH_3	0,000	0,108	0,137	0,193	0,471	N.A.	0	1	1	2	4	9	13	15	25	37	43	355
GA_LOGIC_9	0,000	0,122	0,137	0,242	0,431	0,926	N.A.	1	1	2	4	9	13	15	25	37	43	355
GA_LOGIC_6	0,000	0,102	0,115	0,207	0,368	1,000	1,000	N.A.	0	1	3	8	12	14	24	36	42	354
GA_LOGIC_8	0,000	0,083	0,101	0,198	0,348	1,000	1,000	0,850	N.A.	1	3	8	12	14	24	36	42	354
GA_LOGIC_10	0,000	0,085	0,105	0,180	0,320	0,929	0,905	1,000	1,000	N.A.	2	7	11	13	23	35	41	353
GA_LOGIC_7	0,000	0,078	0,085	0,139	0,291	0,794	0,665	0,795	0,824	0,909	N.A.	5	9	11	21	33	39	351
NN_1LAYER	0,000	0,017	0,014	0,027	0,121	0,421	0,452	0,508	0,519	0,608	0,712	N.A.	4	6	16	28	34	346
GA_LOGIC_5	0,000	0,012	0,015	0,025	0,055	0,294	0,223	0,156	0,156	0,254	0,417	0,789	N.A.	2	12	24	30	342
GA_LOGIC_4	0,000	0,003	0,005	0,013	0,028	0,196	0,180	0,189	0,161	0,228	0,375	0,661	0,913	N.A.	10	22	28	340
GA_LOGIC_3	0,000	0,000	0,000	0,001	0,001	0,035	0,022	0,021	0,021	0,006	0,078	0,208	0,303	0,382	N.A.	12	18	330
nBAYES	0,000	0,000	0,000	0,000	0,000	0,002	0,001	0,002	0,002	0,003	0,006	0,022	0,062	0,078	0,307	N.A.	6	318
GA_LOGIC_2	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,001	0,001	0,001	0,006	0,017	0,034	0,099	0,606	N.A.	312
CENTROIDS	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	N.A.

Este test evalúa la significancia estadística de la diferencia de las tasas de error entre cada uno de los clasificadores. En la tabla, el resultado que se muestra corresponde a cada par de clasificadores multivariables. En el triángulo superior derecho de la tabla se muestra el valor absoluto de la diferencia de las tasas de error para cada par de clasificadores. En el triángulo inferior de la tabla se muestra el p-valor asociado a cada test, para cada par de clasificadores. Un p-valor menor o igual a 0.05 indica que se rechaza la hipótesis nula de igualdad entre ambas tasas de error.

SVM: Support Vector Machines. GA: Algoritmo genético. NN: Perceptrón Multicapa. BAYES_NET: Red Bayesiana.

4.2) Test de Delong

Las curvas ROC para todos los clasificadores multivariados y los mejores cinco clasificadores univariados (Fig 38) muestran que el clasificador generado con SVM es aquél de mejor rendimiento sobre todo el rango posible de umbrales de separación de las clases de modelos correcto e incorrecto.

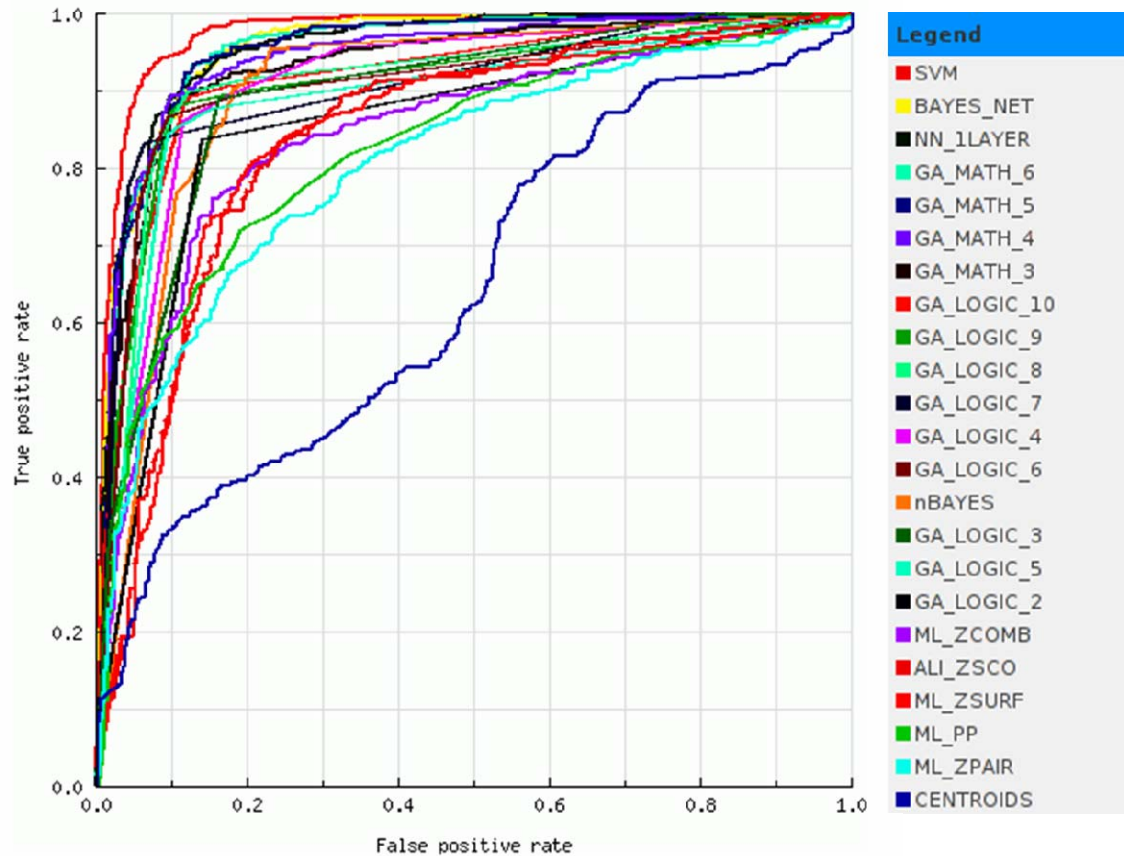


Fig 38. Curvas ROC para los clasificadores multivariados y los mejores cinco clasificadores univariados. La curva correspondiente a SVM está por sobre las otras curvas en todo el rango de falsos positivos versus verdaderos positivos, indicando un mejor rendimiento global con respecto a los otros clasificadores.

Con el fin de determinar si la diferencia entre las áreas de las curvas ROC (AUC) es estadísticamente significativa se aplicó el test de Delong. El resultado para las mejores cinco variables en términos de su AUC (Tabla 13), indica que aquellas variables basadas en propiedades estadísticas son las de mejor rendimiento, junto con la propensión a la partición del modelo (Tabla 23).

Tabla 23. Resultados del Test de Delong para la comparación de clasificadores univariables.

	ML_ZCOMB	ALI_ZSCO	ML_ZSURF	ML_PP	ML_ZPAIR
ML_ZCOMB	N.A.	0.003	0.005	0.021	0.037
ALI_ZSCO	0.793	N.A.	0.002	0.018	0.034
ML_ZSURF	0.379	0.839	N.A.	0.016	0.032
ML_PP	0.165	0.287	0.313	N.A.	0.017
ML_ZPAIR	0.000	0.013	0.005	0.288	N.A.

Este test mide la diferencia de las AUC entre cada una de las mejores cinco variables como clasificadores. En el triángulo superior derecho de la tabla se muestra el valor absoluto de la diferencia de las AUC para cada par de clasificadores. En el triángulo inferior de la tabla se muestra el p-valor asociado a cada test, para cada par de clasificadores. Un p-valor menor o igual a 0.05 indica que se rechaza la hipótesis nula de igualdad entre ambas AUC. Ver Tabla 7 para una descripción del identificador de cada variable.

Los resultados de calcular el intervalo de confianza para cada diferencia de AUC entre cada par de variables como clasificadores se muestra en el Anexo H de este trabajo.

Para el caso de los mejores clasificadores multivariables obtenidos en cuanto a su AUC (Tabla 24), el resultado de medir si la diferencia es estadísticamente significativa indica que SVM es aquél de mejor rendimiento entre los clasificadores generados (Tabla 25).

Tabla 24. AUC obtenido para cada clasificador multivariable.

CLASIFICADOR	AUC
SVM	0.977
BAYES_NET	0.958
NN_1LAYER	0.956
GA_MATH_6	0.954
GA_MATH_5	0.954
NN_2LAYERS	0.950
GA_MATH_4	0.946
GA_MATH_3	0.933
GA_LOGIC_10	0.917
GA_LOGIC_9	0.908
GA_LOGIC_8	0.908
GA_LOGIC_7	0.907
GA_LOGIC_4	0.906
GA_LOGIC_6	0.904
nBAYES	0.900
GA_LOGIC_3	0.891
GA_LOGIC_5	0.888
GA_LOGIC_2	0.850
CENTROIDS	0.645

Los resultados se muestran en orden decreciente. SVM: Support Vector Machines. AUC: área bajo la curva ROC. GA: Algoritmo genético. NN: Perceptrón Multicapa. BAYES_NET: Red Bayesiana.

Tabla 25. Resultados del Test de Delong para la comparación de clasificadores multivariables.

	SVM	BAYES_NET	NN_1LAYER	GA_MATH_6	GA_MATH_5	GA_MATH_4	GA_MATH_3	GA_LOGIC_10	GA_LOGIC_9	GA_LOGIC_8	GA_LOGIC_7	GA_LOGIC_4	GA_LOGIC_6	nBAYES	GA_LOGIC_3	GA_LOGIC_5	GA_LOGIC_2	CENTROIDS
SVM	N.A.	0,019	0,021	0,023	0,023	0,031	0,044	0,060	0,069	0,069	0,069	0,071	0,072	0,077	0,086	0,089	0,127	0,332
BAYES_NET	0,000	N.A.	0,001	0,004	0,004	0,012	0,025	0,041	0,050	0,050	0,050	0,051	0,053	0,058	0,066	0,069	0,108	0,313
NN_1LAYER	0,000	0,779	N.A.	0,003	0,003	0,011	0,023	0,039	0,048	0,049	0,049	0,050	0,052	0,057	0,065	0,068	0,106	0,311
GA_MATH_6	0,000	0,435	0,459	N.A.	0,000	0,008	0,021	0,037	0,046	0,046	0,046	0,047	0,049	0,054	0,062	0,065	0,104	0,309
GA_MATH_5	0,000	0,391	0,460	0,983	N.A.	0,008	0,021	0,037	0,046	0,046	0,046	0,047	0,049	0,054	0,062	0,065	0,103	0,309
GA_MATH_4	0,000	0,032	0,004	0,005	0,028	N.A.	0,013	0,029	0,038	0,038	0,038	0,039	0,041	0,046	0,054	0,057	0,096	0,301
GA_MATH_3	0,000	0,000	0,000	0,000	0,001	0,004	N.A.	0,016	0,025	0,025	0,026	0,027	0,029	0,033	0,042	0,045	0,083	0,288
GA_LOGIC_10	0,000	0,000	0,000	0,000	0,000	0,000	0,048	N.A.	0,009	0,009	0,009	0,011	0,012	0,017	0,026	0,029	0,067	0,272
GA_LOGIC_9	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,164	N.A.	0,000	0,001	0,002	0,004	0,008	0,017	0,020	0,058	0,263
GA_LOGIC_8	0,000	0,000	0,000	0,000	0,000	0,000	0,002	0,196	0,959	N.A.	0,000	0,001	0,003	0,008	0,016	0,019	0,058	0,263
GA_LOGIC_7	0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,210	0,936	0,982	N.A.	0,001	0,003	0,008	0,016	0,019	0,057	0,263
GA_LOGIC_4	0,000	0,000	0,000	0,000	0,000	0,000	0,002	0,191	0,830	0,839	0,901	N.A.	0,002	0,007	0,015	0,018	0,056	0,261
GA_LOGIC_6	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,050	0,568	0,549	0,699	0,821	N.A.	0,005	0,013	0,016	0,054	0,260
nBAYES	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,052	0,338	0,415	0,387	0,495	0,600	N.A.	0,008	0,011	0,050	0,255
GA_LOGIC_3	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,026	0,030	0,063	0,065	0,064	0,333	N.A.	0,003	0,041	0,246
GA_LOGIC_5	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,013	0,000	0,029	0,012	0,006	0,249	0,693	N.A.	0,038	0,243
GA_LOGIC_2	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	N.A.	0,205
CENTROIDS	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	N.A.

Este test evalúa si la diferencia de las AUC entre cada par de clasificadores es estadísticamente significativa. En el triángulo superior derecho de la tabla se muestra el valor absoluto de la diferencia de las AUC para cada par de clasificadores. En el triángulo inferior de la tabla se muestra el p-valor asociado a cada test, para cada par de clasificadores. Un p-valor menor o igual a 0.05 indica que se rechaza la hipótesis nula de igualdad entre ambas AUC.

Los resultados de calcular el intervalo de confianza para cada diferencia de AUC entre dos clasificadores se muestra en el Anexo I de este trabajo.

4.3) Análisis sobre modelos incompletos

La generación de modelos proteicos mediante la técnica de modelado comparativo es particularmente difícil cuando solo una pequeña fracción de la secuencia a modelar – que corresponde a la variable Target Coverage (ALI_tgcov) – es cubierta por la estructura molde. Actualmente, el módulo evaluador del modelado comparativo es tal que el modelo proteico generado para la fracción que sí fue cubierta por el molde es clasificado como incorrecto aún cuando su estructura sea correcta.

Uno de los objetivos de este trabajo es obtener un clasificador que, además de tener un buen rendimiento en general, tenga un buen rendimiento sobre modelos incompletos, esto es, modelos con target coverage (ALI_tgcov) bajo.

En el conjunto de prueba, la presencia de modelos correctos e incorrectos con un bajo target coverage es similar, mientras que la presencia de modelos correctos e incorrectos con un target coverage alto es menor para el primer caso que para el segundo (Fig 39).

Dependiendo del rango de target coverage, los mejores clasificadores en términos de precisión varían. Así por ejemplo, para rangos de target coverage bajos, los clasificadores obtenidos con GA Math muestran un buen rendimiento, mientras que su rendimiento decrece para rangos mayores (Tabla 26).

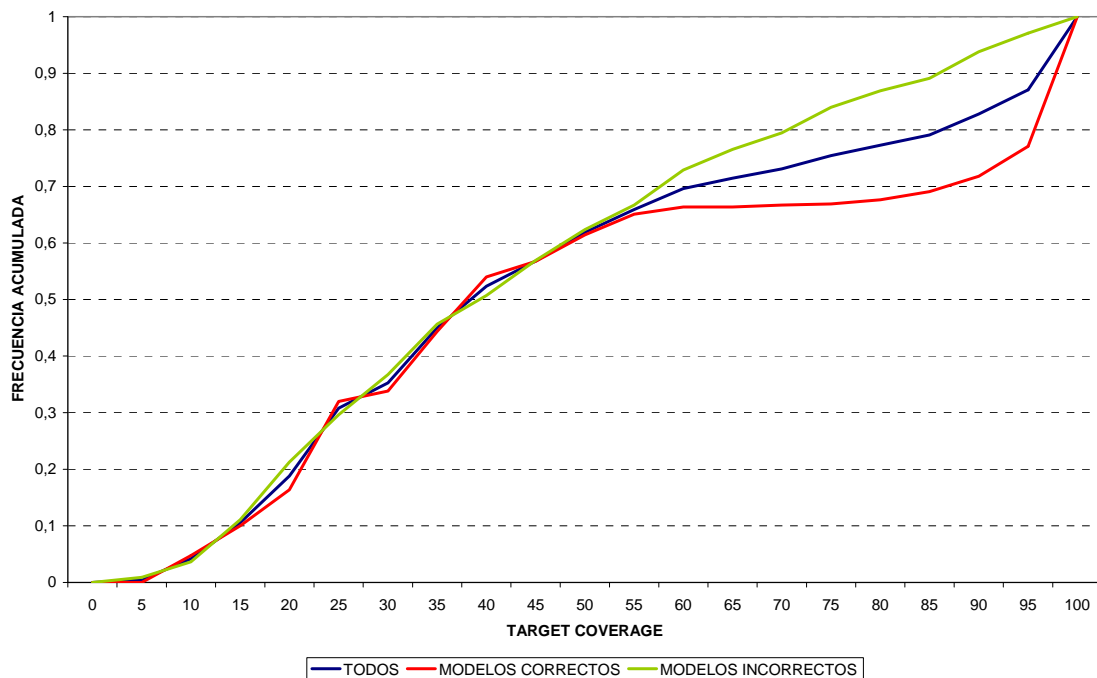


Fig 39. Distribución acumulada de la variable Target Coverage. La distribución acumulada de target coverage se muestra sobre el conjunto de prueba, en los modelos correctos y en los modelos incorrectos que conforman el conjunto de prueba. La mitad de los modelos tienen un target coverage menor o igual al 40%.

Tabla 26. Precisión de los mejores 10 clasificadores sobre los modelos incompletos del conjunto de prueba.

Target Coverage														
0% a 10%			10% a 30%			30% a 50%			50% a 70%			70% a 100%		
CLASIFICADOR	RAZON	ACC	CLASIFICADOR	RAZON	ACC	CLASIFICADOR	RAZON	ACC	CLASIFICADOR	RAZON	ACC	CLASIFICADOR	RAZON	ACC
GA_MATH_3	40/46	0.870	GA_MATH_3	318/342	0.930	SVM	268/293	0.915	SVM	116/123	0.943	SVM	286/296	0.966
GA_MATH_4	40/46	0.870	SVM	314/342	0.918	GA_MATH_5	260/293	0.887	GA_MATH_6	114/123	0.927	BAYES_NET	281/296	0.949
GA_MATH_5	40/46	0.870	GA_MATH_5	310/342	0.906	GA_MATH_4	258/293	0.881	GA_MATH_5	112/123	0.911	GA_LOGIC_6	278/296	0.939
GA_MATH_6	40/46	0.870	GA_LOGIC_8	310/342	0.906	GA_MATH_6	256/293	0.874	GA_MATH_4	109/123	0.886	GA_LOGIC_8	277/296	0.936
TF_COREL	39/46	0.848	BAYES_NET	309/342	0.904	GA_LOGIC_7	252/293	0.860	BAYES_NET	108/123	0.878	GA_LOGIC_3	276/296	0.932
GA_LOGIC_5	39/46	0.848	GA_MATH_4	309/342	0.904	nBAYES	250/293	0.853	GA_LOGIC_6	107/123	0.870	GA_LOGIC_5	276/296	0.932
SVM	39/46	0.848	GA_LOGIC_9	308/342	0.901	GA_LOGIC_6	250/293	0.853	GA_LOGIC_7	107/123	0.870	GA_LOGIC_9	275/296	0.929
GA_LOGIC_4	38/46	0.826	NN_2LAYERS	308/342	0.901	GA_LOGIC_9	249/293	0.850	GA_LOGIC_10	107/123	0.870	GA_MATH_6	274/296	0.926
GA_LOGIC_8	38/46	0.826	GA_MATH_6	307/342	0.898	GA_LOGIC_10	249/293	0.850	NN_1LAYER	107/123	0.870	GA_LOGIC_10	274/296	0.926
GA_LOGIC_10	38/46	0.826	NN_1LAYER	306/342	0.895	BAYES_NET	248/293	0.846	GA_LOGIC_9	105/123	0.854	NN_2LAYERS	274/296	0.926
ML_COREL	37/46	0.804	GA_LOGIC_10	304/342	0.889	NN_1LAYER	248/293	0.846	nBAYES	104/123	0.846	GA_LOGIC_4	273/296	0.922
BAYES_NET	37/46	0.804	GA_LOGIC_7	303/342	0.886	GA_MATH_3	247/293	0.843	GA_MATH_3	103/123	0.837	GA_LOGIC_2	272/296	0.919

Los resultados se muestran para distintos intervalos de la fracción de la secuencia objetivo que fue posible modelar.

5) Discusión

El objetivo principal de este trabajo consistió en la generación de un clasificador óptimo de la calidad de los modelos tridimensionales de proteínas construidos con la técnica de modelado comparativo. Si bien los modelos se separan en clases llamadas correctos e incorrectos, estos modelos corresponden a modelos de mediana calidad (modelos “correctos”) y modelos definitivamente de mala calidad (modelos incorrectos).

Con el fin de entrenar los algoritmos de aprendizaje que generaron los clasificadores a evaluar, se midieron 31 atributos sobre 1153 modelos correctos y 1146 modelos incorrectos. Estos rasgos corresponden a propiedades del alineamiento secuencia-estructura utilizados para construir los modelos, propiedades estructurales del modelo generado y propiedades estructurales del molde utilizado para generar el modelo. Luego, utilizando distintos métodos de ranking, selección y extracción de variables, se midió la relevancia de cada variable con respecto a la clase respuesta y la redundancia entre ellas; esto significó la reducción del número de variables de 31 a 22. El paso siguiente correspondió al entrenamiento de varios algoritmos de aprendizaje con el fin de generar clasificadores que fueron a su vez evaluados sobre un conjunto de prueba compuesto por 550 modelos correctos y 550 modelos incorrectos. Finalmente, aplicando distintas pruebas estadísticas, se midió la significancia de las diferencias en los rendimientos de los clasificadores.

De acuerdo a los resultados obtenidos en cada una de las etapas de este trabajo, se puede plantear la siguiente discusión:

5.1) Análisis estadístico básico de las variables utilizadas.

En general, no se puede asumir un comportamiento normal de la mayoría de los rasgos medidos, a juzgar por los valores de skewness y kurtosis calculados (Tablas 8 y 9) y por los gráficos de normalidad observados para cada uno (Anexo A). Los casos para los cuales sí se podría considerar la hipótesis de normalidad son: ML_zpair para los modelos correctos, y ML_corel, TF_zcomb, TF_zsurf, TF_corel, TW_zcomb y TW_corel para los modelos incorrectos (ver Tabla 7 para una descripción de cada identificador de rasgo). Dado que el criterio de detección de outliers de acuerdo al gráfico de caja y bigotes (Anexo B) se basa en el supuesto de normalidad de los datos, éste no es útil como criterio principal para la eliminación de modelos indeseados en la base de datos inicial. Dado lo anterior, y considerando que se trabajó con un número razonable de variables y de dominio conocido, se decidió descartar aquellos modelos que presentaran cierta suspicacia. En este escenario, se descartaron los modelos incorrectos que presentaban un alto porcentaje de identidad de secuencia en el alineamiento secuencia-estructura ($ALI_seqide > 40\%$), debido principalmente a que éstos constituyen casos de proteínas que han sido resueltas bajo diferentes condiciones experimentales que influyen de manera relevante en su conformación final (Ej. Las calmodulinas en presencia o ausencia de calcio sufren cambios conformacionales de gran envergadura). Esto permitió descartar 40 modelos incorrectos. No se encontraron casos considerados como outliers en los modelos correctos.

5.2) *Ranking de las variables.*

Los resultados obtenidos con las tres medidas utilizadas – grado de relación funcional, razón de entropía e incertidumbre simétrica - para el ranking de las variables de acuerdo a su relevancia con respecto a la clase respuesta son en general coherentes entre sí (Tabla 10 y Fig. 31). Las cinco mejores propiedades al momento de considerar su relevancia corresponden a cuatro propiedades estadísticas (ML_zpair, ML_zsurf, ML_zcomb, ALI_zsco) y una propiedad física (ML_pp). Estas propiedades ya han sido descritas como relevantes para la evaluación de la calidad de modelos proteicos en estudios anteriores [40] con otro conjunto de datos que incluía modelos de muy buena calidad y modelos con un alto grado de error.

Tres de ellas, el Z-score de la superficie accesible (ML_zsurf), Z-score de los pares residuo-residuo (ML_zpair) y Z-score combinado (ML_zcomb) ya han sido reportados como los mejores predictores de la calidad de los modelos [42]; su alta relevancia es esperable dado que los potenciales estadísticos son herramientas que capturan gran parte de la complejidad estructural presente en las estructuras proteicas nativas [111]. En el caso de los potenciales de pares residuo-residuo (ML_zpair), éstos capturan la propensión a la interacción de cada par posible de residuos en base a su distancia espacial. Por otra parte, los potenciales de superficie accesible (ML_zsurf) capturan la propensión a la interacción de los residuos con el solvente. El potencial combinado (ML_zcomb) resulta de la suma normalizada de los potenciales de pares residuo-residuo y potenciales de superficie accesible. Cabe recordar que los potenciales estadísticos que se calculan para medir estos Z-scores asignan un valor positivo a aquellas instancias del criterio utilizado para definirlo que se observan con baja frecuencia y un valor negativo a aquellas más frecuentes. En este caso, los modelos incorrectos tienen en promedio un valor mayor que los modelos correctos, lo que indica que sobre estos últimos se encuentran en mayor proporción interacciones esperables por naturaleza que en el caso de los modelos incorrectos. Estas mismas propiedades medidas sobre el molde utilizado para generar el modelo – ya sea sobre el molde completo o sobre la fracción efectivamente utilizada para la generación del modelo – TF_zpair, TF_zsurf, TF_zcomb y TW_zpair, TW_zsurf y TW_zcomb son menos relevantes para la clase respuesta que cuando éstas se miden a partir de los modelos, lo que indica una distribución más sobrepuesta de sus valores en ambas clases de modelos proteicos.

El Z-score del alineamiento secuencia-estructura (ALI_zsco) es otra propiedad estadística relevante para la clase respuesta. Esta propiedad está basada en la matriz de sustitución de residuos que se utiliza en el alineamiento. La matriz de sustitución provee, para cada par de aminoácidos, un puntaje que indica si la sustitución de un aminoácido por otro en un par de secuencias proteicas es observada con frecuencia en proteínas relacionadas o pertenecientes a una misma familia. Así, esta propiedad mide la calidad del alineamiento en términos de cuan similares pueden ser dos residuos alineados, aunque éstos no sean exactamente residuos idénticos, y por lo tanto, junto con su naturaleza estadística, es capaz de capturar más información que la simple identidad del alineamiento (ALI_seqide), lo que explica su mayor relevancia con respecto a la clase respuesta.

La propiedad física que resultó ser más relevante para la clase respuesta fue aquella de la propensión a la partición del modelo (ML_pp). Los modelos incorrectos tienen en promedio una mayor propensión a la partición que los modelos correctos (2.22 versus

1.86), lo que indica que estos últimos tienden a tener su región hidrofóbica inmersa en mayor proporción dentro de la estructura en comparación a los modelos incorrectos. Esta misma propiedad medida sobre la región efectiva del molde utilizado para generar el modelo y sobre el molde completo (TF_pp y TW_pp respectivamente) tiene un promedio mucho más cercano entre ambas clases (1.88 versus 1.80 y 1.88 versus 1.83 respectivamente) y distribuciones más superpuestas, lo que explica su menor relevancia para la clase respuesta.

Las propiedades de Z-score de superficie accesible, Z-score de pares residuo-residuo, Z-score combinado y propensión a la partición medidas sobre el modelo proteico son las únicas para las cuales la correlación con las mismas propiedades medidas sobre el molde utilizado para generar los modelos es relativamente baja. Para todos los otros atributos medidos sobre las categorías correspondientes a los modelos, el molde completo, y la fracción de molde efectivo utilizado para generar el modelo, la correlación es extremadamente alta, lo que explica su posicionamiento cercano en el ranking. Todas las propiedades medidas sobre la estructura del molde completo utilizado para generar el modelo están altamente correlacionadas con las mismas propiedades medidas sobre el fragmento del molde efectivamente utilizado para generar el modelo.

5.3) Selección de variables

El objetivo de aplicar algoritmos de selección de variables es obtener un subconjunto del conjunto inicial de variables que contenga aquellas que sean las más relevantes para la clase respuesta y las menos redundantes entre sí.

El algoritmo de Yu y Liu, que utiliza la incertidumbre simétrica como medida de relevancia y redundancia, toma la variable que considera más relevante en cada iteración, y elimina aquellas variables que son más redundantes que relevantes para la clase respuesta. En este caso, el algoritmo se ejecutó para un total de cuatro iteraciones. En la primera iteración, la variable correspondiente al Z-score de energía de la superficie accesible (ML_zsurf) fue considerada como la más relevante y 19 variables (entre ellas todas las propiedades estadísticas) se eliminan por ser consideradas redundantes. Esto demuestra el significativo poder de discriminación que tiene esta variable por sí sola. En la segunda iteración, la propensión a la partición del modelo (ML_pp) es la variable más relevante, pero no hay variables que sean más redundantes que ella, y por lo tanto ninguna variable es eliminada. En la tercera iteración, el orden de contacto relativo del modelo (ML_pp) es aquella variable más relevante, eliminando propiedades geométricas tales como el radio de giro y compactación. En la cuarta y final iteración, la identidad de secuencia del alineamiento (ALI_seqide) es aquella más relevante. Así, el conjunto final propuesto por el algoritmo de Yu y Liu como aquél que maximiza la relevancia y minimiza la redundancia es aquél compuesto por: ML_corel, ML_pp y ML_zsurf y ALI_seqide. Las tres primeras variables presentan una buena relevancia con respecto a la clase respuesta, como se puede apreciar en los resultados del ranking. La variable ALI_seqide, si bien tiene una menor relevancia por sí sola, este resultado indica que contiene información única, esto es, no reemplazable por otras variables. Si bien uno pudiese esperar que ALI_zsco sea la variable elegida en lo que refiere a las propiedades del alineamiento secuencia-estructura, el hecho de haber sido descartada previamente por la variable ML_zsurf indica que hay cierta redundancia

entre ambas variables y por lo tanto la única variable que contiene información referente al alineamiento y que debe ser preservada es ALI_seqide.

El algoritmo de Koller y Sahami, a diferencia del algoritmo de Yu y Liu, no provee un subconjunto final de variables, sino que el usuario debe decidir un número de variables a considerar como aquél que conformará el subconjunto final. El concepto clave detrás de este algoritmo, el manto de Markov, corresponde a un conjunto de variables que “cubre” la información de una variable cualquiera (que, por supuesto, no pertenece al manto), y que por lo tanto puede ser eliminada. Así, el tamaño del manto de Markov (K), afecta de manera directa el resultado (Tabla 12). En este trabajo se probaron tres valores distintos del manto de Markov: $K=0$, $K=1$ y $K=2$. El caso $K=0$ es equivalente a no considerar la redundancia entre las variables al momento de medir su relevancia para la clase respuesta, por lo tanto, para este valor, el algoritmo de Koller y Sahami se transforma en un método de ranking más; los resultados de este algoritmo para $K=0$ son coherentes con los resultados obtenidos con los tres métodos de ranking. El tamaño óptimo del manto de Markov depende de la naturaleza del conjunto de variables en sí. Un manto de Markov de tamaño uno, esto es, una variable cubre la información de otra variable candidata a ser eliminada, puede no ser suficiente si las variables no están lo suficientemente correlacionadas entre sí. Por lo tanto, se considera prudente utilizar un manto de Markov de tamaño dos ($K=2$) como un parámetro que asegura preservar la información de las variables eliminadas en el proceso de selección. Para $K=2$, si se toman las primeras cinco variables (ML_zsurf, ALI_seqide, TW_len, ML_comp y ML_pp), tres de ellas, ML_zsurf, ALI_seqide y ML_pp están también presentes en el subconjunto óptimo que entrega el algoritmo de Yu y Liu.

5.4) Análisis de Componentes Principales

El método de análisis de componentes principales busca capturar la información presente en las 31 variables estudiadas en un número de variables menor (las componentes principales), no correlacionadas entre sí, que resultan de la combinación lineal de las variables originales. El criterio que comúnmente se aplica para definir el número de componentes principales que resumen la mayor parte de la información corresponde a aquellas que capturan un 90% de la variabilidad de las variables originales. Idealmente, las primeras dos componentes principales podrían lograr este objetivo, pero en este caso las primeras dos componentes sólo capturan el 59.1% de la información. Para capturar el 90% de variabilidad con las componentes generadas tendrían que tomarse las primeras 9 componentes principales. Si bien esto permite en principio reducir el número de variables a utilizar en la siguiente fase de entrenamiento de algoritmos de aprendizaje de 31 a 9, el hecho de que sean nueve variables no facilita el análisis en términos de su visualización, y más importante aún, se pierde el dominio, en términos de interpretación, existente sobre las variables originales. La ponderación de cada variable original sobre las primeras cuatro componentes principales (Fig 32 y Fig 33) evidencia que la redundancia de la mayoría de las variables corresponden a la misma propiedad medida sobre el modelo y el molde utilizado.

Luego de los resultados obtenidos para el análisis estadístico básico de las variables, la aplicación de los métodos de ranking, selección y extracción de variables, se concluye

que el conjunto de variables correspondiente a las propiedades medidas sobre el molde completo utilizado para generar el modelo proteico son altamente redundantes con respecto a las mismas propiedades medidas sobre la región efectiva del molde utilizado para generar el modelo proteico. Además, son poco relevantes con respecto a la clase respuesta. Esto permitió descartarlas del análisis, reduciendo así el número de variables a trabajar de 31 a 22, compuesto por las propiedades del alineamiento secuencia-estructura, propiedades del modelo generado y propiedades de la región efectiva del molde utilizado para generar los modelos proteicos. Si bien se puede argumentar que la variable TW_len fue seleccionada por el algoritmo de Koller y Sahami dentro del subconjunto de las más relevantes, la alta correlación de esta variable con sus equivalentes en el modelo (ML_len, $r = 0.965$) y en la fracción del molde utilizado para generar el modelo (TF_len $r = 0.985$) permite pensar que la exclusión de TW_len no se traduce en una pérdida significativa de información.

Así, y tras uniformar el tamaño de cada clase a 1100 modelos cada una, el conjunto de datos con el que se contó para la siguiente etapa de generación y evaluación de clasificadores correspondió a 1100 modelos correctos, 1100 modelos incorrectos, y 22 variables medidas sobre ellos.

En las siguientes secciones se provee la discusión correspondiente a la generación y evaluación de diferentes clasificadores uni y multivariados. Cabe señalar que, si bien los resultados en términos de rendimiento (precisión y área bajo la curva ROC, o AUC) que se utilizaron para la comparación de dos clasificadores corresponden a aquellos medidos sobre el conjunto de prueba, también se reportaron los rendimientos medidos sobre el conjunto de entrenamiento como una forma de control del sobreajuste de cada clasificador al momento de ser generado con el algoritmo de aprendizaje correspondiente.

La discusión que se provee a continuación respecto a los clasificadores tiene la siguiente estructura. Primero, se discuten los resultados obtenidos para cada variable considerada como un clasificador; esto permite entender el rendimiento que los clasificadores multivariados deben sobrepasar para ser útiles al problema planteado. Segundo, se provee una discusión parcial para cada algoritmo de aprendizaje por separado, hasta el punto en que se obtiene, para cada uno de ellos, el clasificador óptimo que es evaluado sobre el conjunto de prueba. Tercero, se discute de manera global los rendimientos obtenidos para los diferentes clasificadores óptimos para cada tipo de algoritmo de aprendizaje, basándose en los ensayos de significancia estadística empleados. Por último, basándose en lo anterior, se declara a uno de ellos como el clasificador óptimo para el problema de clasificación planteado.

5.5) Variables como clasificadores

Con el fin de medir cuan efectiva es cada una de las variables como clasificador de la calidad de los modelos proteicos, se midió su rendimiento en términos de precisión (o accuracy) y área bajo la curva ROC (AUC). Esto permitió tener una idea de cuan competitivas son las variables con respecto a los clasificadores multivariados generados. En términos de precisión, los resultados obtenidos (Tabla 13 y Tabla 20) indican que las variables que mejor actúan como clasificadores son (en orden

descendente de precisión): (i) el Z-score combinado del modelo (ML_zcomb), (ii) el Z-score de superficie accesible del modelo (ML_zsurf), (iii) el Z-score del alineamiento secuencia estructura (ALI_zsco), (iv) la propensión a la partición del modelo (ML_pp) y (v) el orden de contacto relativo del modelo (ML_corel). El test estadístico de McNemar para las diferencias de la precisión entre dos clasificadores muestra que, si bien el mejor clasificador es ML_zcomb, su diferencia en precisión no es estadísticamente significativa con respecto a aquella para ML_zsurf y ALI_zsco; la diferencia sí es significativa con ML_pp y ML_corel.

En términos de AUC, los resultados obtenidos (Tabla 13 y Tabla 23) indican que las variables que mejor actúan como clasificadores son: (i) el Z-score combinado del modelo (ML_zcomb), (ii) el Z-score del alineamiento secuencia estructura (ALI_zsco), (iii) el Z-score de superficie accesible del modelo (ML_zsurf), (iv) la propensión a la partición del modelo (ML_pp) y (v) el Z-score de los pares residuo-residuo del modelo (ML_zpair). Como era de esperar de acuerdo a la relevancia de cada variable para la clase respuesta, las propiedades estadísticas son las mejores en términos de poder de discriminación, seguidas de la propensión a la partición. Cabe señalar que era esperable que ML_zcomb fuera la mejor propiedad estadística, pues combina información (no redundante) de las interacciones intramoleculares de la estructura como de la interacción de cada uno de los aminoácidos con el solvente. El ensayo estadístico de Delong para la diferencia de AUC entre dos clasificadores muestra que, si bien el mejor clasificador es ML_zcomb, su diferencia en AUC no es estadísticamente significativa con respecto a aquella para ALI_zsco, ML_zsurf y ML_pp; la diferencia sí es significativa con ML_zpair.

Así, se concluye que las variables con mayor poder de discriminación son ML_zcomb, ML_zsurf y ALI_zsco. La máxima precisión y AUC obtenido corresponde a aquellas obtenidas para ML_zcomb, con 80.2% de precisión y 0.846 de AUC. Esto determina el mínimo rendimiento que deben tener los clasificadores multivariados para ser considerados como una alternativa útil al problema de clasificación en cuestión en este trabajo.

5.6) Clasificadores multivariados

Con el fin de generar una alternativa efectiva a los clasificadores representados por cada una de las propiedades medidas sobre los modelos proteicos, diferentes algoritmos de aprendizaje se entrenaron para obtener clasificadores que involucren dos o más de las 22 propiedades a considerar tras el análisis de ranking, selección y extracción de variables.

A continuación se entrega una breve discusión por cada algoritmo de aprendizaje utilizado y clasificador generado, para luego concluir sobre el total de los clasificadores, de manera de definir a uno de ellos como el óptimo dentro de las alternativas generadas.

5.6.1) Clasificador basado en la distancia a centroides

El clasificador basado en la distancia a centroides es el más simple dentro de la clase de clasificadores multivariados. Para el total de modelos en el conjunto de entrenamiento,

para cada clase (correctos o incorrectos), se estima el centroide con el promedio aritmético, que corresponde en este caso a un vector en \mathcal{R}^{22} (Tabla 14). Luego, cada modelo en el conjunto de prueba se asocia a la clase cuyo promedio esté más cercano.

El bajo rendimiento obtenido con este clasificador tanto en términos de precisión (56.3%) como de AUC (0.645) sugiere que las nubes de puntos de los modelos correspondientes a ambas clases están fuertemente sobrelapadas, lo cual sugiere a su vez que este problema de clasificación no es linealmente separable. Así, el resultado obtenido demuestra que, si bien este clasificador es conceptualmente simple, no es útil para el problema de clasificación enfrentado en este trabajo, pero sirve de control para entender la complejidad del mismo. Así, este clasificador resulta en un punto de referencia respecto al cual se espera que todos los otros métodos de clasificación multivariantes tengan un rendimiento significativamente mayor.

5.6.2) Algoritmo genético GA Math

El algoritmo genético GA Math evoluciona fórmulas matemáticas que sirven de clasificadores sobre los modelos proteicos. Las fórmulas resultantes pueden ser combinaciones lineales o no lineales de las variables originales. El parámetro que hace la diferencia entre obtener fórmulas más simples o más complejas corresponde a la profundidad máxima del árbol que representa a la fórmula matemática; valores más pequeños de este parámetro resultan en fórmulas más simples, y valores más grandes resultan en fórmulas más complejas. Con el fin de generar fórmulas de distinta complejidad, se utilizaron cuatro valores para este parámetro: 3, 4, 5 y 6. Cabe mencionar que, independiente de la profundidad del árbol, los parámetros correspondientes al tamaño de la población y al número de iteraciones para alcanzar un óptimo se fijaron en valores relativamente bajos tras observar que valores más altos de estos no generaban clasificadores significativamente mejores.

Las fórmulas óptimas obtenidas para cada profundidad de árbol se observan en la sección 3.6.3.1 de los resultados. Cada fórmula matemática fue llamada GA_Math_X, donde X corresponde a la profundidad del árbol utilizado para generar esa fórmula ($X=\{3,4,5,6\}$). Estas corresponden, para cada profundidad, a la mejor fórmula en términos de rendimiento sobre el conjunto de prueba, obtenida sobre 100,000 ejecuciones independientes.

Dentro de las cuatro fórmulas óptimas, aquellas con profundidad 5 y 6 (GA_Math_5 y GA_Math_6) son las que exhiben un mejor rendimiento sobre el conjunto de prueba (Tabla 15), lo que es esperable pues son fórmulas más complejas. Sin embargo, cuando se compara el rendimiento en términos de la precisión, el test de McNemar (Tabla 22) indica que la diferencia del GA_Math_6 con los otros GA_Math no es estadísticamente significativa ($p\text{-valor}>0.05$). Cuando se compara el rendimiento en términos del AUC, el test de DeLong (Tabla 25) indica que la diferencia del GA_Math_6 con GA_Math_5 no es estadísticamente significativa, pero sí lo es al momento de evaluarse su diferencia con los otros dos clasificadores, GA_Math_4 y GA_Math_3.

La frecuencia de aparición de las distintas variables en las 100,000 ejecuciones para cada profundidad (Fig 35) sugiere que aquellas de mayor relevancia para la conformación de fórmulas óptimas serían el Z-score de energía combinado del modelo

(*ML_zcomb*), el Z-score del alineamiento secuencia-estructura (*ALI_zsco*), el Z-score de la superficie accesible del modelo (*ML_zsurf*), el orden de contacto relativo del fragmento (*TF_corel*) y el orden de contacto relativo del modelo (*ML_corel*). Todas estas variables son buenos clasificadores por sí solos, por lo que se espera su aparición en las fórmulas óptimas. Una variable de buen rendimiento por sí sola, la propensión a la partición del modelo (*ML_pp*), aparece con baja frecuencia, y sin embargo variables que tienen un rendimiento bajo, como lo son la compactación del modelo (*ML_comp*) y el radio de giro del modelo (*ML_rg*) aparecen con una frecuencia significativa. Esto muestra que la combinación lineal o no lineal de variables para generar un clasificador no necesariamente involucra sólo aquellas de mejor rendimiento, obteniéndose combinaciones que tienen poder discriminante pero que no son evidentes.

Es interesante observar que hay ciertas estructuras en las fórmulas óptimas que se repiten. Por ejemplo, $ML_PP \cdot (ALI_ZSCO + ML_RG)$ es un factor que se repite en tres de las cuatro fórmulas obtenidas. También, $ML_RG \cdot ML_ZCOMB$ es una estructura que aparece ponderada por el cuadrado de una variable que representa el orden de contacto relativo del modelo (*ML_corel* para el caso de *GA_Math_5*) o del fragmento de molde utilizado para generar el modelo (*TF_corel* para el caso de *GA_Math_6*), dos variables que están muy correlacionadas ($r = 0.986$). Esto, junto con el poco incremento del rendimiento al pasar de las fórmulas *GA_Math_5* a *GA_Math_6*, demuestra que basta con considerar una profundidad máxima de 5 para obtener un clasificador óptimo dentro de esta clase de algoritmo de aprendizaje, para el problema enfrentado en este trabajo.

5.6.3) Algoritmo genético GA logic

El algoritmo genético GA Logic evoluciona las ramas de un árbol de decisión que sirve de clasificador sobre los modelos proteicos. Dada la naturaleza continua de las variables utilizadas en este trabajo, las fórmulas lógicas resultantes corresponden a reglas que particionan el espacio de las variables originales involucradas utilizando operadores de desigualdad, unidas por “y” (^) lógico. Luego, cada rama es conectada por el operador “o” (v) lógico. El parámetro que hace la diferencia entre obtener fórmulas lógicas más simples o más complejas corresponde al número de genes que representa el número de variables involucradas en la fórmula lógica; valores más pequeños de este parámetro resultan en fórmulas más simples, y valores más grandes resultan en fórmulas más complejas. Con el fin de generar fórmulas de distinta complejidad, se utilizaron nueve valores de este parámetro: desde 2 hasta 10. Cabe mencionar que, independiente del número de genes, los parámetros correspondientes al tamaño de la población y al número de iteraciones para alcanzar un óptimo se fijaron en valores relativamente bajos tras observar que valores más altos de estos no generaban clasificadores significativamente mejores.

Las fórmulas lógicas óptimas obtenidas para cada número de genes se observan en la sección 3.6.3.2 de los resultados. Cada fórmula lógica ha sido llamada *GA_Logic_X*, donde X corresponde al número de genes utilizado para generarla ($X = \{2, \dots, 10\}$). Estas corresponden, para cada número de genes, a la mejor fórmula lógica en términos de rendimiento sobre el conjunto de prueba, obtenida sobre 100,000 ejecuciones independientes.

Dentro de las nueve fórmulas óptimas, aquella generada con un número de genes igual a 9 (GA_Logic_9) es la que exhibe el mejor rendimiento sobre el conjunto de prueba en términos de precisión, mientras que GA_Logic_10 es la que exhibe el mejor rendimiento en términos de AUC (Tabla 16), lo que es esperable pues son las fórmulas lógicas más complejas. Sin embargo, cuando se compara el rendimiento en términos de la precisión, el test de McNemar (Tabla 22) indica que la diferencia de precisión del GA_Logic_9 con los otros clasificadores generados con este método no es estadísticamente significativa, a excepción de GA_Logic_2 y GA_Logic_3 (p-valor<0.05). Cuando se compara el rendimiento en términos del AUC, el test de DeLong (Tabla 25) indica que la diferencia del GA_Logic_10 con los demás GA Logic sólo es significativa cuando se compara con GA_Logic_2, GA_Logic_3 y GA_Logic_5 (p-valor<0.05).

La frecuencia de aparición de las distintas variables en las 100,000 ejecuciones para cada profundidad (Fig 36) indica que aquellas de mayor relevancia para la conformación de fórmulas lógicas óptimas son el Z-score de la superficie accesible del modelo (ML_zsurf), el radio de giro del fragmento de molde utilizado para generar el modelo (TF_rg), seguidas del el Z-score del alineamiento secuencia-estructura (ALI_zsco), el PSI-BLAST e-value del alineamiento (ALI_pbeval) (aunque esta dos no tienen incidencia en las fórmulas lógicas generadas con 2 genes) y el radio de giro del modelo (ML_rg). Es esperable que ML_zsurf y ALI_zsco estén dentro de las variables más frecuentes, dada su relevancia con respecto a la clase respuesta. Sin embargo, una variable de buen rendimiento por sí solo, como lo es la propensión a la partición del modelo (ML_pp), aparece con baja frecuencia, mientras que variables que tienen un rendimiento bajo, como lo son TF_rg y ML_rg aparecen con una frecuencia significativa. Esto muestra que la partición del espacio que proveen las fórmulas lógicas para generar un clasificador no necesariamente involucra sólo aquellas de mejor rendimiento, obteniéndose fórmulas lógicas que tienen poder discriminante pero que no son evidentes.

En las dos secciones anteriores se han discutido los resultados de los clasificadores generados con algoritmos genéticos que evolucionan fórmulas matemáticas (GA Math) y ramas de árboles de decisión (GA Logic). Hay aspectos en común entre ambos algoritmos de aprendizaje que se discuten a continuación:

Primero, si bien se pudo realizar una etapa de optimización de parámetros para cada instancia de este algoritmo de aprendizaje, se decidió no hacerlo dado el alto número de parámetros involucrados. Los valores utilizados se decidieron en base a ejecuciones previas del algoritmo genético, y no se espera un conjunto de parámetros distintos que permita obtener rendimientos significativamente mayores. Segundo, respecto a los parámetros correspondientes al número de genes y profundidad de árbol utilizados para generar cada clasificador GA Logic y GA Math respectivamente, en principio se podría pensar, dado que son un parámetro más del algoritmo genético, que debió utilizarse sólo un valor en lugar de considerar nueve valores distintos para generar nueve clasificadores distintos en el caso del GA Logic, o cuatro valores distintos en el caso del GA Math. Sin embargo, dado que estos parámetros inciden directamente en la complejidad del clasificador generado, necesariamente valores mayores resultan en clasificadores de mejor rendimiento, y por lo tanto la comparación de sus rendimientos en la fase de entrenamiento no sería justa. Esto, junto con el hecho de que lo que queremos es

obtener, además de un clasificador de rendimiento óptimo, un clasificador lo más simple posible, llevó a decidir utilizar todos los clasificadores generados para los distintos valores de número de genes y profundidad de árbol descritos previamente. Tercero, valores muy altos de estos parámetros, como ya se mencionó para los casos GA_Logic_9, GA_Logic_6, GA_Math_5 y GA_Math_6 resultan en una fórmula más compleja y no necesariamente en un mejor clasificador. Cuarto, si bien es esperable que el rendimiento de un clasificador sobre el conjunto de entrenamiento sea mayor que sobre el conjunto de prueba, la pequeña diferencia de rendimiento observada tanto en términos de precisión como de AUC entre estos dos conjuntos para todas las instancias de los clasificadores generados con los dos algoritmos genéticos, demuestra que no hay un sobreajuste significativo sobre el conjunto de entrenamiento. Finalmente, cabe destacar el alto poder discriminante de estos clasificadores, considerando que no utilizan todas las variables disponibles y que proveen clasificadores que permiten generar entendimiento humano detrás del proceso de clasificación (se sabe exactamente cómo participa cada variable), al contrario de otros clasificadores que se transforman en verdaderas cajas negras, y que por lo tanto cualquier resultado de clasificación no puede ser sometido a mayor análisis.

5.6.4) Naive Bayes

El clasificador que proporciona Naive Bayes es uno de los más simples conceptualmente, pues clasifica a cada modelo de acuerdo a la clase que es más probable, y porque asume independencia de las variables. Este clasificador corresponde a un caso particular de la red Bayesiana, en el cual el nodo correspondiente a la clase respuesta referencia de manera directa a cada una de las variables, las cuales no tienen conexión alguna entre sí. Para calcular la probabilidad de cada instancia de la clase respuesta (modelos correctos e incorrectos) es necesario estimar la probabilidad asociada a cada variable; esta estimación se puede llevar cabo de dos maneras: (i) asumiendo una distribución normal de las variables y estimando a su vez el promedio y varianza, o (ii) discretizando el rango de valores de cada variable. Los resultados obtenidos sobre el conjunto de entrenamiento para las dos modalidades de estimación de la probabilidad de las variables indican que es mejor en términos de rendimiento discretizar las variables que asumir una distribución normal. Si bien la diferencia en la precisión promedio para ambos casos no es grande, es esperable que asumir una distribución normal tenga un peor rendimiento, dado los resultados obtenidos en el análisis básico de las variables para ambas clases, en el cual se tiene que, salvo excepciones, las variables no siguen una distribución normal.

Al aplicar el clasificador optimizado (esto es, estimando las probabilidades por medio de discretización) sobre el conjunto de prueba, se obtiene que el rendimiento es de 85.18% en precisión y 0.900 en AUC, lo que corresponde a un rendimiento no satisfactorio, pero que era esperable dada la simplicidad del clasificador y el supuesto de independencia que claramente no se aplica para muchas de las variables.

5.6.5) Perceptrón multicapa

El perceptrón multicapa corresponde a un algoritmo de aprendizaje que se basa en un problema de optimización sin restricciones, en el cual se busca minimizar el error representado por la diferencia entre la salida real y la salida esperada de la red (representada por neuronas conectadas en las capas de entrada, ocultas y de salida). La

minimización del error se logra, para cada ciclo de aprendizaje, por medio del ajuste de los pesos de las conexiones entre las neuronas. A diferencia de las máquinas de vectores de soporte, este algoritmo no se detiene por sí sólo, por lo que es necesario establecer un criterio de detención del entrenamiento de manera de evitar un sobreajuste de la red a los datos de entrenamiento. En este caso, el criterio utilizado correspondió al rendimiento del clasificador generado en cada ciclo de aprendizaje sobre un conjunto independiente al de entrenamiento, de manera que cuando el error sobre este conjunto comenzara a crecer, entonces se detuviera el aprendizaje. Este algoritmo de aprendizaje involucra cuatro parámetros que debieron ser optimizados: (i) Parámetros propios de la arquitectura de la red, correspondiente al número de neuronas por cada capa oculta y al número de capas ocultas propiamente tal, y (ii) Parámetros propios del problema de optimización, correspondientes a la tasa de aprendizaje α y al momento η . Los resultados obtenidos en la fase de entrenamiento indican que aquel perceptrón multicapa óptimo dentro de esta clase corresponde a aquél con una capa oculta, doce neuronas en la capa oculta, $\alpha = 0.1$ y momento $\eta=0.0$ (Tabla 17). Esto indica que el aprendizaje de la red se genera a una velocidad más bien lenta, y el efecto de considerar el momento para controlar la velocidad de convergencia (lo que está proporcionado por η) es despreciable.

El rendimiento sobre el conjunto de prueba de esta red optimizada resultó en 87.73% de precisión y 0.956 de AUC, lo cual no es satisfactorio considerando la naturaleza compleja de este algoritmo de aprendizaje y el hecho que utiliza las 22 variables para la generación del clasificador.

5.6.6) Redes Bayesianas

Las redes Bayesianas, como su nombre lo indica, corresponden a un algoritmo de aprendizaje que se basa en la regla de Bayes para obtener aquella estructura de un grafo acíclico dirigido que mejor se ajusta a los datos proporcionados en el conjunto de entrenamiento. Una vez definida la estructura de grafo óptima, se estima la probabilidad de cada nodo (esto es, cada variable) dado los padres. Si bien este algoritmo de aprendizaje se puede aplicar a muchos problemas de aprendizaje supervisado y no supervisado, en el caso de los problemas de clasificación la restricción existente sobre la estructura de la red es que aquel nodo que representa a la clase respuesta debe ir en la “raíz” del grafo, de manera de poder determinar la probabilidad de cada instancia de la clase respuesta dadas las variables (en este caso, las distintas propiedades medidas sobre los modelos proteicos). La optimización del grafo se llevó a cabo por medio de la utilización de diferentes algoritmos de búsqueda sobre el espacio de posibles arquitecturas. El resultado obtenido es que el grafo generado con el algoritmo de búsqueda TAN es el óptimo dentro de los posibles generados para este algoritmo de aprendizaje. Recordemos que una de las ventajas de utilizar las redes Bayesianas es que permiten visualizar las dependencias de las variables a través de la estructura del grafo. En este caso, la estructura resultante con el algoritmo TAN (estructura no mostrada) evidencia dependencias que ya eran conocidas de la etapa de ranking, selección y extracción de variables. Así por ejemplo, los nodos correspondientes al radio de giro (TF_rg), orden de contacto relativo (TF_corel), compactación (TF_comp) y largo (TF_len) del fragmento efectivo de molde utilizado para generar las variables son padres de los nodos que representan a las respectivas variables medidas sobre los modelos (ML_rg, ML_corel, ML_comp y ML_len respectivamente), cada par de las

cuales está fuertemente correlacionado entre sí. Así también, variables no tan correlacionadas entre sí, como lo son la propensión a la partición del modelo (ML_pp) y del fragmento de molde utilizado para generarlo (TF_pp) no tienen una dependencia directa de acuerdo al grafo.

El resultado de utilizar el grafo óptimo como clasificador sobre los modelos pertenecientes al conjunto de prueba entrega una precisión de 89.36% y un AUC de 0.958, lo cual es un rendimiento satisfactorio dada la simplicidad del clasificador. Cabe mencionar que un factor que puede afectar al rendimiento de este clasificador es el hecho de que las variables son discretizadas, lo que necesariamente se traduce en una pérdida de información al momento de realizarse la clasificación de cada modelo.

5.6.7) Máquinas de vectores de soporte

Las máquinas de vectores de soporte corresponden a un tipo de algoritmo de aprendizaje que plantea un problema de optimización que genera una solución única, correspondiente a un hiperplano (o hipersuperficie) separador entre los modelos pertenecientes a las clases correcto e incorrecto. Como se establece en la sección 2.7.4, si bien en principio el problema de optimización se plantea para obtener soluciones a problemas de clasificación linealmente separables, la incorporación de las funciones núcleo permite extender esta herramienta para problemas que no son linealmente separables. Además, la incorporación de un término en el problema de optimización que tolera cierto grado de error en la clasificación, permite la generación de una solución más general.

Para el problema de clasificación enfrentado en este trabajo, dado que las distribuciones de las variables y el resultado del clasificador basado en distancia a centroides indican que el problema es no linealmente separable, y dado que es deseable una solución genérica para futuros modelos a clasificar, se decidió utilizar este algoritmo de aprendizaje incluyendo tanto funciones núcleo como el término adicional en el problema de optimización que permite una holgura en cuanto al error cometido. Esto implicó una serie de parámetros que debieron ser optimizados para generar un clasificador óptimo dentro de los posibles para este algoritmo de aprendizaje (Tabla 2 y Tabla 3). Con respecto a las funciones núcleo, aquella que permite obtener una mayor precisión corresponde a una función Gaussiana con un factor de ponderación γ de la diferencia entre dos puntos \mathbf{x} e \mathbf{y} igual a 0.9 (Tabla 19). El parámetro C corresponde al costo que implica considerar variables de holgura en el problema de optimización; un valor pequeño de este parámetro significa tolerar varias “holguras” (Fig 20), mientras que un valor grande significa tolerar pocas “holguras” distintas de cero. Diferentes valores de este parámetro tienen poca incidencia en el rendimiento de los clasificadores generados durante la fase de entrenamiento y optimización de parámetros (Anexo E), siendo $C=3.6$ para la función núcleo Gaussiana aquella que maximiza la precisión promedio (94.82 %).

Así, el clasificador óptimo dentro de este tipo de algoritmo corresponde a un clasificador generado con una función núcleo Gaussiana con $\gamma = 0.9$ y con un costo $C=3.6$ (Anexo F). La precisión obtenida sobre el conjunto de prueba del clasificador generado con estos parámetros es 93.0%, mientras que el AUC corresponde a 0.977, lo cual corresponde a un rendimiento satisfactorio para el problema en cuestión.

5.7) Comparación de clasificadores

Hasta ahora se ha discutido acerca de los clasificadores uni y multivariados enfocados sólo en la variable o algoritmo de aprendizaje que genera el clasificador, respectivamente. Dado que el objetivo de este trabajo es declarar uno de los clasificadores como óptimo dentro de todos los clasificadores considerados, es necesario enfocarse en la comparación global del rendimiento de los mismos. No olvidemos que la hipótesis con la cual se trabajó es que múltiples variables relacionadas al alineamiento secuencia-estructura, al modelo proteico generado y al molde utilizado para generar el modelo, proporcionan la información necesaria para generar clasificadores que sobrepasan de manera estadísticamente significativa al rendimiento de cada una de las variables por sí sola, incluidas aquellas propiedades basadas en potenciales estadísticos que han sido reportadas previamente como buenos predictores de la calidad de un modelo proteico.

El rendimiento de los clasificadores fue evaluado con dos métricas: la precisión, correspondiente al porcentaje de modelos correctamente clasificados, y el área bajo la curva ROC (AUC).

Para comparar el rendimiento de los clasificadores en base a la precisión, se utilizó el test estadístico de McNemar. Este ensayo, de carácter no paramétrico, genera un estadístico chi-cuadrado que permite calcular un p-valor que proporciona la evidencia necesaria para decidir si dos clasificadores cualesquiera tienen una tasa de error igual o distinta. Los resultados del test de McNemar sobre los mejores clasificadores univariados (Tabla 20) muestran que el mejor clasificador univariado es el Z-score combinado del modelo (ML_zcomb), con una precisión de 80.2%. Cuando se compara este clasificador con los clasificadores multivariados generados, estos últimos son significativamente superiores en rendimiento (p -valor <0.05), excepto por el clasificador basado en distancia a centroides, en cuyo caso el clasificador univariado lo supera significativamente en rendimiento (p -valor <0.05). Dentro de los clasificadores multivariados, aquel de mayor precisión es SVM (93.0%). Este rendimiento es significativamente superior a todos los otros clasificadores multivariados (p -valor <0.05 , Tabla 22). Así, de acuerdo al test de McNemar, el mejor clasificador es aquel generado con SVM.

Para comparar el rendimiento de los clasificadores en base al AUC, se utilizó el test estadístico de DeLong. Este test, de carácter no paramétrico, genera un estadístico chi-cuadrado que permite calcular un p-valor que proporciona la evidencia necesaria para decidir si dos clasificadores cualesquiera tienen un AUC igual o distinta. Los resultados del test de DeLong sobre los mejores clasificadores univariados (Tabla 23) muestran que el mejor clasificador univariado es el Z-score combinado del modelo (ML_zcomb), con un AUC de 0.846. Cuando se compara este clasificador con los clasificadores multivariados generados, estos últimos son significativamente superiores en rendimiento (p -valor <0.05), excepto por el clasificador basado en distancia a centroides, en cuyo caso el clasificador univariado supera significativamente en rendimiento (p -valor <0.05), y el clasificador GA_Logic_2, en cuyo caso la diferencia no es estadísticamente significativa (p -valor >0.05). Dentro de los clasificadores multivariados, aquel de mayor AUC es SVM, con un 0.977. Este rendimiento es significativamente superior a todos los otros clasificadores multivariados (p -valor <0.05 , Tabla 25). Además, su rendimiento es superior para todo el rango de

especificidad versus sensibilidad (Fig 38). Así, de acuerdo al test de Delong, el mejor clasificador es aquel generado con SVM.

El hecho de que el algoritmo de aprendizaje SVM genere el clasificador óptimo para el problema planteado en este trabajo es esperable dado su fácil extensión a problemas de clasificación no linealmente separables, así como la inclusión de términos en el problema de optimización que permiten obtener soluciones más generales dada la población de individuos.

Así, se tiene que la hipótesis de este trabajo se valida, pues existen clasificadores generados con múltiples variables que sobrepasan en rendimiento a aquellas variables por si solas consideradas como clasificadores. Cabe notar que la exploración de una amplia gama de clasificadores fue esencial para la correcta demostración de la hipótesis, pues si, por ejemplo, se hubiera considerado sólo el clasificador basado en distancia a centroides, se tendría el resultado opuesto.

Dado la discusión anterior, y considerando el conjunto de datos utilizado en este trabajo, se declara como el clasificador óptimo para la evaluación de la calidad de los modelos proteicos, al clasificador generado con el algoritmo de aprendizaje Support Vector Machine (SVM), con una función de kernel gaussiana con factor de ponderación de 0.9 y un costo $C=3.6$.

5.8) Consideración del espacio de modelos incompletos

Un obstáculo común en el modelado comparativo al momento de discriminar correctamente si un modelo proteico es correcto o incorrecto tiene relación con la proporción de la secuencia objetivo que se pudo efectivamente modelar, esto es, la variable Target Coverage (ALI_tgcov).

Dada una secuencia a modelar y dado el alineamiento del molde con una baja fracción de la secuencia objetivo, si bien la pequeña fracción cubierta puede estar correctamente modelada, actualmente el modulo de evaluación descarta con alta probabilidad el modelo, pues su naturaleza estructural “parcial” se contradice con lo que el módulo evaluador considera es una buena estructura en términos de las interacciones presentes (especialmente en las interacciones de la proteína con el solvente que la rodea, dado que en los modelos proteicos incompletos no se está representando de buena forma la realidad). Así, un objetivo adicional de este trabajo consistió en la determinación de un clasificador que resolviera este obstáculo de manera satisfactoria.

La proporción de modelos correctamente clasificados para distintos intervalos de target coverage (Tabla 26) muestra que el clasificador declarado como óptimo, SVM, tiene un rendimiento satisfactorio para esta clase de modelos. Cabe señalar, que para valores de target coverage extremadamente bajos (<30%) los clasificadores GA Math representan una alternativa de igual eficiencia sobre esta clase de modelos, pero con una diferencia respecto a SVM en lo que se refiere al número de modelos incompletos correctamente clasificados muy baja. Así, el clasificador GA Math resulta un clasificador satisfactorio sobre esta clase de modelos.

6) Conclusiones

A partir de los resultados obtenidos y de la discusión planteada, se puede concluir lo siguiente:

- Las variables medidas sobre el alineamiento secuencia-estructura en el modelado comparativo, así como las variables geométricas, físicas y estadísticas medidas sobre los modelos proteicos generados y el molde utilizado para generar los modelos proveen información útil para la correcta clasificación de la calidad de los modelos.
- Las propiedades geométricas, físicas y estadísticas medidas sobre el molde completo utilizado para generar el modelo son altamente redundantes en información con las mismas propiedades medidas sobre la fracción del molde efectivamente utilizado para generar el modelo.
- Clasificadores univariados basados en el Z-score de energía combinado del modelo, Z-score de energía de pares residuo-residuo del modelo, y Z-score de energía de superficie accesible del modelo son aquellos con mayor poder de discriminación en la evaluación de la calidad de los modelos proteicos.
- La inclusión de propiedades del alineamiento secuencia-estructura, del modelo generado y del molde utilizado para generar el modelo, permiten generar clasificadores multivariados que son significativamente mejores en rendimiento que los mejores clasificadores univariados utilizados en este trabajo, lográndose un aumento en precisión del 13% y un aumento en área bajo la curva ROC de 0.131.
- El clasificador generado con el algoritmo de aprendizaje Support Vector Machine, con kernel gaussiano, con $\gamma = 0.9$ y con un costo $C=3.6$ es aquél de mejor rendimiento tanto en términos de precisión como de área bajo la curva ROC, y por lo tanto se declara como el clasificador óptimo para la evaluación de la calidad de modelos tridimensionales de proteínas. Además, su rendimiento global en lo que se refiere al rango posible de especificidad versus sensibilidad es mayor que aquel de los otros clasificadores uni y multivariados.
- El clasificador óptimo generado en este trabajo es útil para el problema de la correcta evaluación de modelos proteicos incompletos. Adicionalmente, para modelos proteicos con una fracción modelada extremadamente baja (<30%), los clasificadores generados con el algoritmo genético GA_Math representan una alternativa equivalente o igualmente útil.

7) Referencias

1. Darwin, C., J. Murray, and William Clowes and Sons., *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life*. 1859, London: John Murray ... : Printed by W. Clowes and Sons ... ix, [1], 502 p., [1] folded leaf of plates.
2. Darwin, C., *The origin of species*. 1909, New York,: P. F. Collier & son. 2 p. l., 3-553, [1] p.
3. Anfinsen, C.B., et al., *The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain*. Proc Natl Acad Sci U S A, 1961. **47**: p. 1309-14.
4. Anfinsen, C.B., *Principles that govern the folding of protein chains*. Science, 1973. **181**(96): p. 223-30.
5. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins*. Embo J, 1986. **5**(4): p. 823-6.
6. Chothia, C., *Proteins. One thousand families for the molecular biologist*. Nature, 1992. **357**(6379): p. 543-4.
7. Holm, L. and C. Sander, *Globin fold in a bacterial toxin*. Nature, 1993. **361**(6410): p. 309.
8. Holm, L. and C. Sander, *The FSSP database of structurally aligned protein fold families*. Nucleic Acids Res, 1994. **22**(17): p. 3600-9.
9. Holm, L. and C. Sander, *Structural similarity of plant chitinase and lysozymes from animals and phage. An evolutionary connection*. FEBS Lett, 1994. **340**(1-2): p. 129-32.
10. Holm, L. and C. Sander, *Evolutionary link between glycogen phosphorylase and a DNA modifying enzyme*. Embo J, 1995. **14**(7): p. 1287-93.
11. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures*. J Mol Biol, 1977. **112**(3): p. 535-42.
12. Berman, H.M., et al., *The Protein Data Bank*. Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.
13. Baker, D. and A. Sali, *Protein structure prediction and structural genomics*. Science, 2001. **294**(5540): p. 93-6.
14. Sanchez, R. and A. Sali, *Large-scale protein structure modeling of the Saccharomyces cerevisiae genome*. Proc Natl Acad Sci U S A, 1998. **95**(23): p. 13597-602.
15. Guex, N., A. Diemand, and M.C. Peitsch, *Protein modelling for all*. Trends Biochem Sci, 1999. **24**(9): p. 364-7.
16. Jones, D.T., *GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences*. J Mol Biol, 1999. **287**(4): p. 797-815.
17. Kihara, D., et al., *Ab initio protein structure prediction on a genomic scale: application to the Mycoplasma genitalium genome*. Proc Natl Acad Sci U S A, 2002. **99**(9): p. 5993-8.
18. Schwede, T., et al., *SWISS-MODEL: An automated protein homology-modeling server*. Nucleic Acids Res, 2003. **31**(13): p. 3381-5.
19. Lesk, A.M., *CASP2: report on ab initio predictions*. Proteins, 1997. **Suppl 1**: p. 151-66.

20. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure*. Science, 1991. **253**(5016): p. 164-70.
21. Godzik, A., A. Kolinski, and J. Skolnick, *Topology fingerprint approach to the inverse protein folding problem*. J Mol Biol, 1992. **227**(1): p. 227-38.
22. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A new approach to protein fold recognition*. Nature, 1992. **358**(6381): p. 86-9.
23. Sippl, M.J. and S. Weitckus, *Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations*. Proteins, 1992. **13**(3): p. 258-71.
24. Fiser, A., et al., *Comparative protein structure modeling*. Computational Biochemistry and Biophysics, 2001: p. 275-312.
25. Marti-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes*. Annu Rev Biophys Biomol Struct, 2000. **29**: p. 291-325.
26. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. J Mol Biol, 1970. **48**(3): p. 443-53.
27. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
28. Greer, J., *Comparative model-building of the mammalian serine proteases*. J Mol Biol, 1981. **153**(4): p. 1027-42.
29. Levitt, M., *Accurate modeling of protein conformation by automatic segment matching*. J Mol Biol, 1992. **226**(2): p. 507-33.
30. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints*. J Mol Biol, 1993. **234**(3): p. 779-815.
31. Sippl, M.J., *Knowledge-based potentials for proteins*. Curr Opin Struct Biol, 1995. **5**(2): p. 229-35.
32. Lazaridis, T. and M. Karplus, *Effective energy function for proteins in solution*. Proteins, 1999. **35**(2): p. 133-52.
33. Haykin, S. and S. Haykin, *Neural Networks: A comprehensive foundation*. 2nd ed. 1998: Prentice Hall.
34. Kohonen, T., *Self-organizing maps*. 1997, New York: Springer-Verlag.
35. Kohonen, T., *Self-organization and Associative Memory*. 1998, New York: Springer-Verlag.
36. Goldberg, D., *Genetic Algorithms in Search, Optimization, and Machine Learning*. 1989: Addison-Wesley.
37. Jolliffe, I., *Principal Component Analysis*. 2nd ed. 2002: Springer-Verlag.
38. Mitchell, T., *Machine Learning*. 1997, McGraw-Hill. p. 52-78.
39. Cristianini, N. and J. Shawe-Taylor, *An introduction to Support Vector Machines and other kernel-based learning methods*. 1st ed. 2000: Cambridge University Press.
40. Melo, F. and A. Sali, *Fold Assessment for Comparative Protein Structure Modelling*. 2007. **16**: p. 2412-2426.
41. McGuffin, L.J. and D.T. Jones, *Improvement of the GenTHREADER method for genomic fold recognition*. Bioinformatics, 2003. **19**(7): p. 874-81.
42. Melo, F., R. Sanchez, and A. Sali, *Statistical potentials for fold assessment*. Protein Sci, 2002. **11**(2): p. 430-48.
43. Sippl, M.J., *Recognition of errors in three-dimensional structures of proteins*. Proteins, 1993. **17**(4): p. 355-62.

44. Lemer, C.M., M.J. Rooman, and S.J. Wodak, *Protein structure prediction by threading methods: evaluation of current techniques*. Proteins, 1995. **23**(3): p. 337-55.
45. Park, B.H., E.S. Huang, and M. Levitt, *Factors affecting the ability of energy functions to discriminate correct from incorrect folds*. J Mol Biol, 1997. **266**(4): p. 831-46.
46. Melo, F. and E. Feytmans, *Novel knowledge-based mean force potential at atomic level*. J Mol Biol, 1997. **267**(1): p. 207-22.
47. Melo, F. and E. Feytmans, *Assessing protein structures with a non-local atomic interaction energy*. J Mol Biol, 1998. **277**(5): p. 1141-52.
48. Melo, F., et al., *ANOLEA: a www server to assess protein structures*. Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 187-90.
49. Melo, F. and M.A. Marti-Renom, *Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets*. Proteins, 2006. **63**(4): p. 986-95.
50. Sippl, M.J., *Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures*. J Comput Aided Mol Des, 1993. **7**(4): p. 473-501.
51. Sippl, M.J., *Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins*. J Mol Biol, 1990. **213**(4): p. 859-83.
52. Bondi, A., *Van der Waals volumes and radii*. J.Phys.Chem., 1964. **68**: p. 441-451.
53. Thomas, P.D. and K.A. Dill, *Statistical potentials extracted from protein structures: how accurate are they?* J Mol Biol, 1996. **257**(2): p. 457-69.
54. Bonneau, R., et al., *Contact order and ab initio protein structure prediction*. Protein Sci, 2002. **11**(8): p. 1937-44.
55. Baker, D., *A surprising simplicity to protein folding*. Nature, 2000. **405**(6782): p. 39-42.
56. Chambers, J., et al., *Graphical methods for data analysis*. 1983: Wadsworth.
57. Shannon, C., *A mathematical theory of communication*. The Bell System Technical Journal, 1948. **27**: p. 379-423.
58. Kullback, S., *Information Theory and Statistics*. 1959: General Publishing Company.
59. Press, W., et al., *Numerical recipes in C: the art of scientific computing*. 2nd ed. 1992: Cambridge University Press.
60. Li, W., *Mutual Information Functions versus Correlation Functions*. Journal of Statistical Physics, 1990. **60**(5/6): p. 823-837.
61. Blum, A. and P. Langley, *Selection of relevant features and examples in machine learning*. Artificial intelligence, 1997. **97**(1-2): p. 245-271.
62. Pearl, J., *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. 1984: Addison-Wesley.
63. Biesiada, J. and W. Duch, *Feature Selection for High-Dimensional Data: A Kolmogorov-Smirnov Correlation-Based Filter*.
64. Bins, J. and B. Draper, *Feature Selection from Huge Feature Sets*. Proceedings of International Conference in Computer Vision, 2001. **2**: p. 159-165.
65. Cantu-Paz, E., *Feature Subset Selection By Estimation Of Distribution Algorithms*. Proceedings of the Genetic and Evolutionary Computation Conference, 2002: p. 303 - 310.

66. Cantu-Paz, E., *Feature subset selection, class separability, and genetic algorithms*. Genetic and Evolutionary Computation Conference, 2004: p. 959-970.
67. Chen, Y. and C. Lin (2005) *Combining SVMs with various feature selection strategies*. **Volume**,
68. Duch, W., et al., *Feature Selection based on Information Theory, Consistency and Separability Indices*. Proceedings of the 9th International Conference on Neural Information Processing, 2002. **4**: p. 1951-1955.
69. Fleuret, F., *Fast Binary Feature Selection with Conditional Mutual Information*. Journal of Machine Learning Research, 2004. **5**: p. 1531-1555.
70. Gilad-Bachrach, R., A. Navot, and N. Tishby, *Margin Based Feature Selection - Theory and Algorithms*. Proceedings of the 21st International Conference on Machine Learning, 2004: p. 337-344.
71. Hall, M. and L. Smith, *Practical Feature Subset Selection for Machine Learning*. Proceedings of the Australian Computer Science Conference, 1996.
72. Lanzi, P., *Fast feature selection with genetic algorithms: a filter approach*. IEEE International Conference on Evolutionary Computation, 1997: p. 537-540.
73. Liu, H., et al., *Evolving Feature Selection*. IEEE Intelligent Systems and Their Applications, 2005(6): p. 64-76.
74. Liu, H. and R. Setiono, *A Probabilistic Approach to Feature Selection - A Filter Solution*. Proceedings of International Conference of Machine Learning, 1996: p. 319-327.
75. Ruiz, R., J. Aguilar-Ruiz, and J. Riquelme, *SOAP: Efficient Feature Selection of Numeric Attributes*. Lecture Notes in Computer Science, 2002(2527): p. 233-242.
76. Vafaie, H. and K. DeJong, *Genetic Algorithms as a Tool for Feature Selection in Machine Learning*. Proceedings of the 1992 IEEE Int. Conf. on Tools with AI, 1992: p. 200-203.
77. Wang, H., D. Bell, and F. Murtagh, *Axiomatic Approach to Feature Subset Selection Based on Relevance*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999. **21**(3): p. 271-276.
78. Yang, J. and V. Honavar, *Feature Subset Selection using a Genetic Algorithm*. IEEE intelligent systems & their applications, 1998. **13**(2): p. 44-49.
79. Radijovac, P., et al., *Feature Selection Filters based on the Permutation Test*. European conference on machine learning, 2004. **3201**: p. 334-346.
80. Richeldi, M. and P. Lanzi, *ADHOC: a tool for performing effective feature selection*. Tools with Artificial Intelligence, 1996., Proceedings Eighth IEEE International Conference on, 1996: p. 102-105.
81. Krishnapuram, B., et al., *A Bayesian approach to joint feature selection and classifier design*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. **26**(9): p. 1105-1111.
82. Liu, H. and R. Setiono, *Chi2: feature selection and discretization of numeric attributes*. Tools with Artificial Intelligence, 1995. Proceedings., Seventh International Conference on, 1995: p. 388-391.
83. Koller, D. and M. Sahami, *Toward Optimal Feature Selection*. In: Proc. of the ICML, 1996: p. 129-134.
84. Kullback, S. and R. Leibler, *On information and sufficiency*. Annals of Mathematical Statistics, 1951. **22**: p. 76-86.
85. Yu, L. and H. Liu, *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*. ICML-03, 2003. **20**(2): p. 856-863.

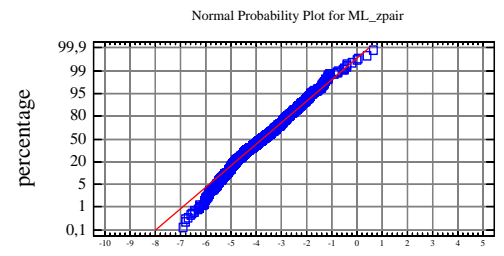
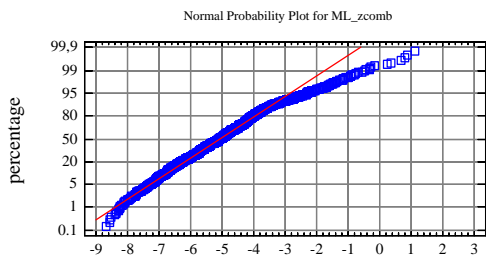
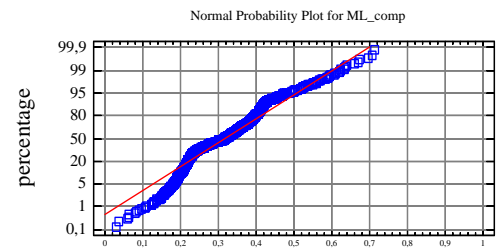
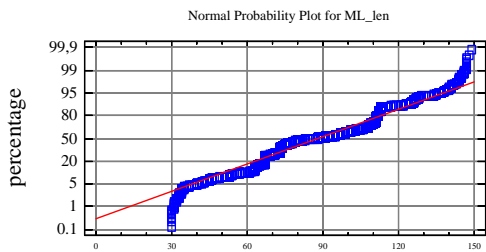
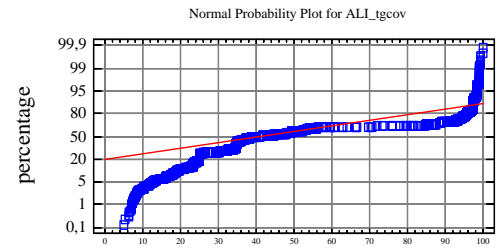
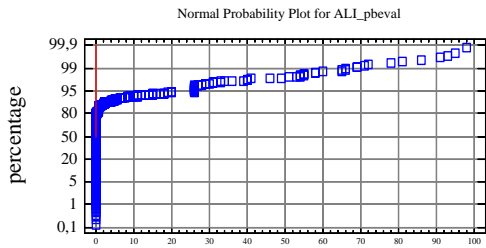
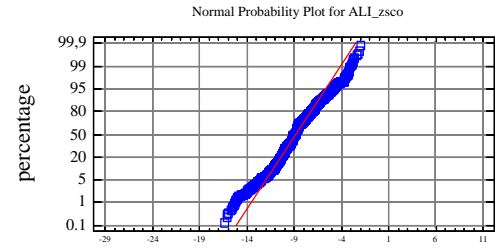
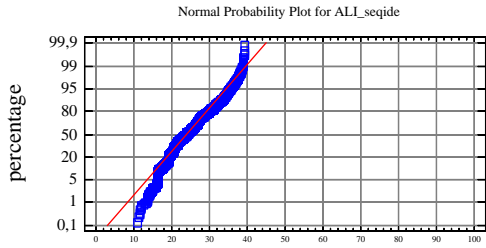
86. Fausett, L., *Fundamentals of Neural Networks*. Prentice Hall, 1994.
87. Hyvärinen, A., J. Karhunen, and E. Oja, *Independent Component Analysis*. 2001: John Wiley and Sons.
88. Holland, J., *Adaptation in Natural and Artificial Systems*. 1975: MIT Press.
89. Dietterich, T.G., *Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*. *Neural Comput*, 1998. **10**(7): p. 1895-1923.
90. Vapnik, V., *The nature of statistical learning theory*. 1995, New York: Springer-Verlag.
91. Vapnik, V., *Statistical learning theory*. 1998: Wiley.
92. Vapnik, V., *An overview of statistical learning theory*. *IEEE transactions on neural networks*, 1999. **10**(5): p. 988-999.
93. Hernandez, J., M. Ramirez, and C. Ferri, *Introducción a la Minería de Datos*. 2005: Prentice Hall.
94. Isasi, P. and I. Galván, *Redes de Neuronas Artificiales: un enfoque práctico*. 2004: Prentice Hall.
95. Rosenblatt, F., *The perceptron: a perceiving and recognizing automation*. Technical Report 85-460-1, 1957.
96. Rosenblatt, F., *The perceptron: a theory of statistical separability in cognitive systems*. Technical Report VG-1196-G-1, 1958.
97. Minsky, M. and S. Papert, *Perceptrons: an introduction to computational geometry*. 1969: MIT Press.
98. Rumelhart, D., G. Hinton, and R. Williams, *Parallel distributed processing, chapter: Learning representation by backpropagating errors*. 1986: MIT Press.
99. Cooper, G. and E. Herskovits, *A bayesian method for the induction of probabilistic networks from data*. *Machine Learning*, 1992. **9**: p. 309-347.
100. Buntine, W.L., *A guide to the literature on learning probabilistic networks from data*. *IEEE Transactions on Knowledge and Data Engineering*, 1996. **8**: p. 195-210.
101. Cheng, J. and R. Greiner, *Comparing bayesian network classifiers*. *Proceedings UAI*, 1999: p. 101-107.
102. Friedman, N., D. Geiger, and M. Goldszmidt, *Bayesian Network Classifiers*. *Machine Learning*, 1997. **29**: p. 131-163.
103. Chow, C.K. and C.N. Liu, *Approximating discrete probability distributions with dependence trees*. *IEEE Trans on Info. Theory*, 1968. **IT-14**: p. 426-467.
104. Salzberg, S.L., *On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach*. *Data Mining and Knowledge Discovery*, 1997. **1**(3): p. 317-328.
105. Fawcett, T., *ROC graphs: notes and practical considerations for researchers*. *Kluwer Academic Publisher*, 2004. **1**: p. 1-38.
106. Swets, J.A., *Measuring the accuracy of diagnostic systems*. *Science*, 1988. **240**: p. 1285-1293.
107. Swets, J.A., R.M. Dawes, and J. Monahan, *Better decisions through science*. *Scientific American*, 2000. **283**: p. 82-87.
108. Ling, C., J. Huang, and H. Zhang, *AUC: a statistically consistent and more discriminating measure than accuracy*. *IJCAI-03*, 2003.
109. Everitt, B.S., *The analysis of contingency tables*. *Chapman and Hall*, London, 1977.
110. Delong, E.R., D.M. Delong, and D.L. Clarke-Pearson, *Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach*. *Biometrics*, 1988. **44**(3): p. 837-845.

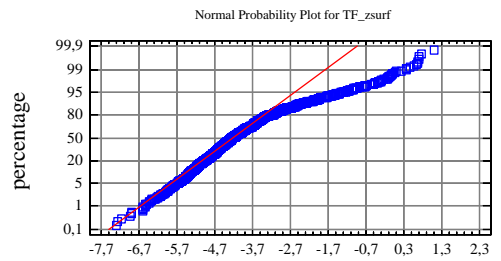
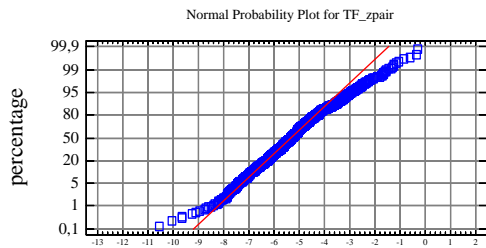
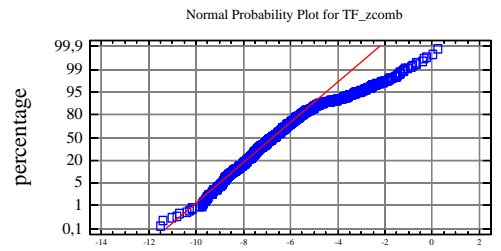
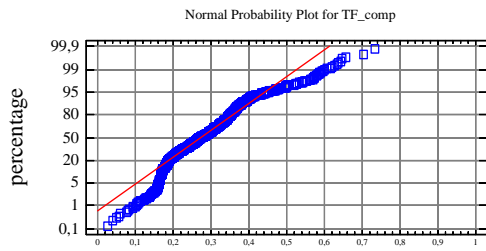
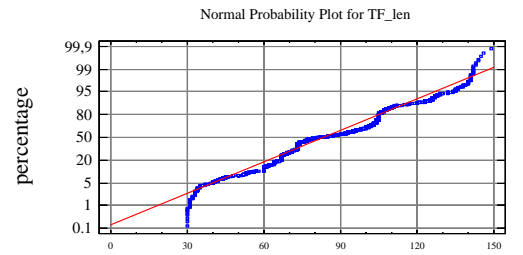
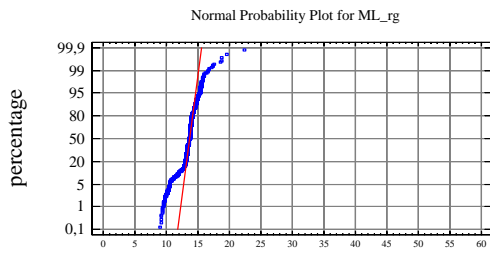
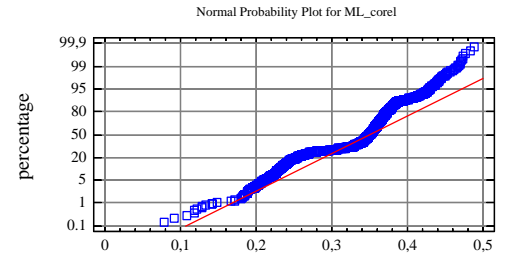
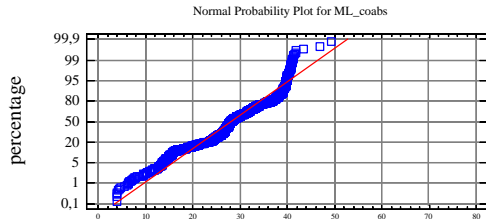
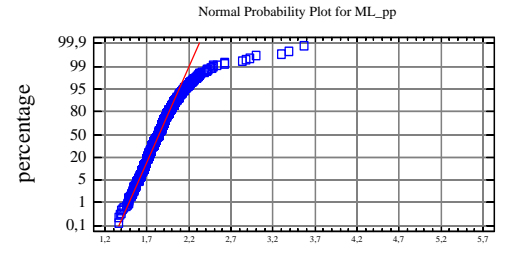
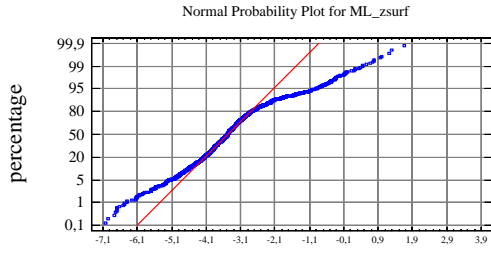
111. Melo, F. and E. Feytmans, *Scoring functions for protein structure prediction*, in *Computational Structural Biology*, M.P.a.T. Schwede, Editor. 2008.

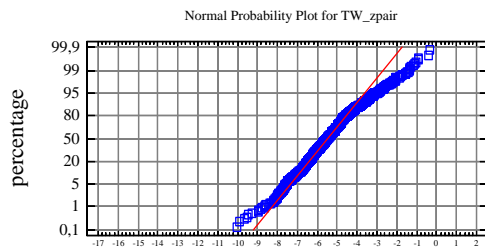
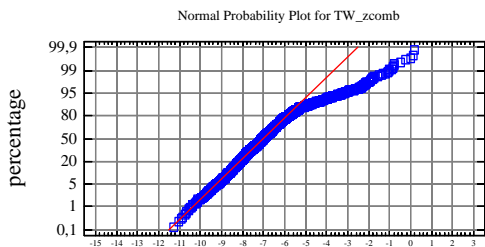
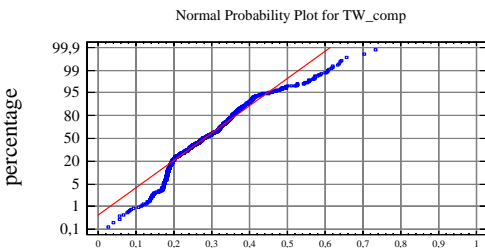
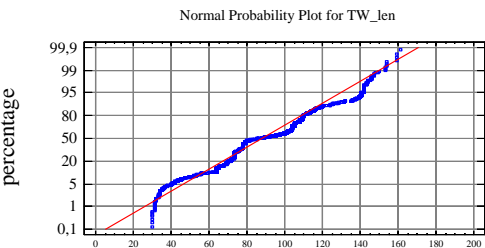
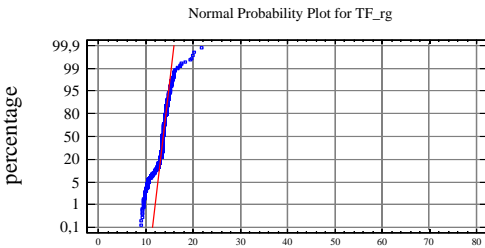
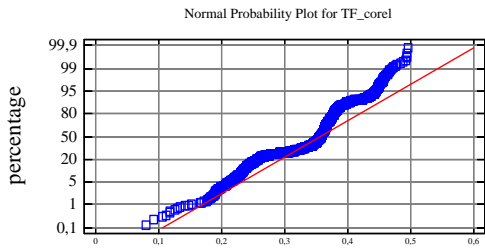
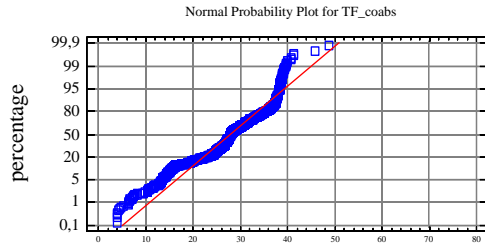
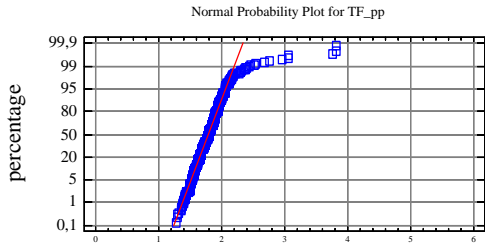
8) Anexos

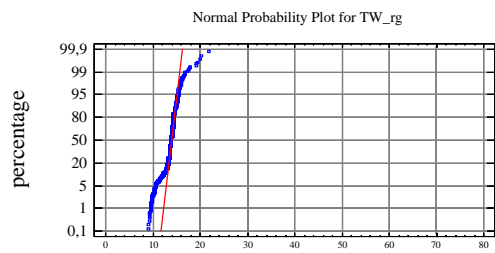
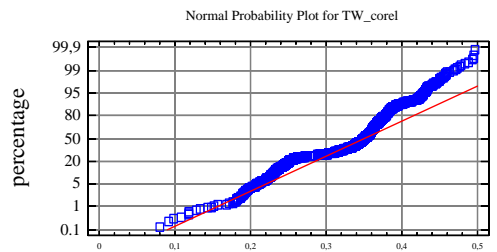
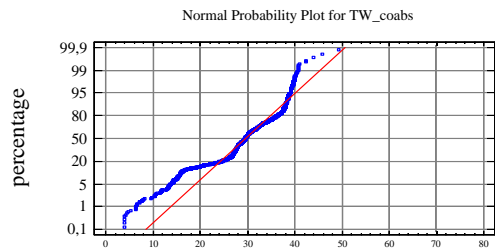
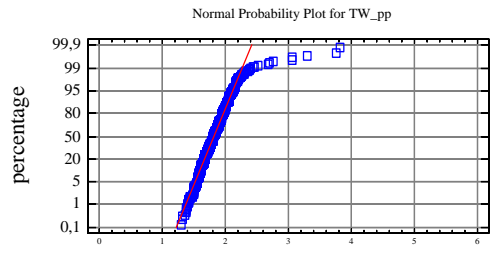
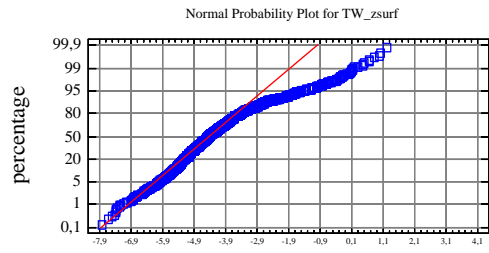
Anexo A: gráficos de probabilidad normal

A continuación se presentan los gráficos de probabilidad normal (normal probability plots) sobre los modelos correctos:

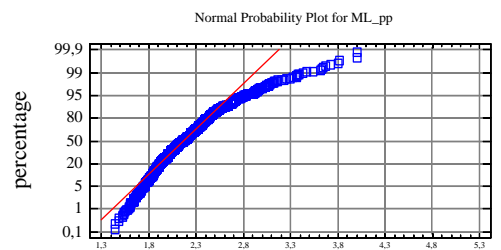
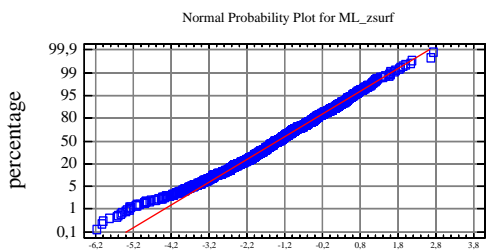
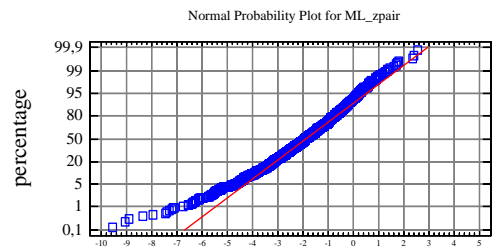
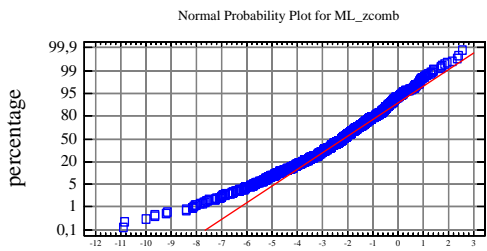
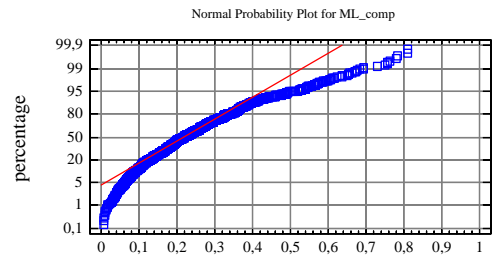
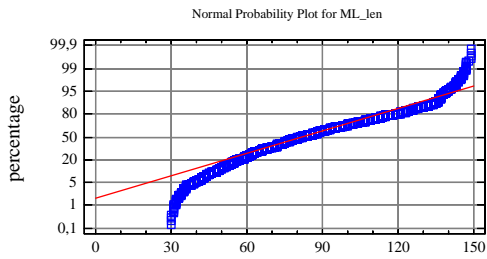
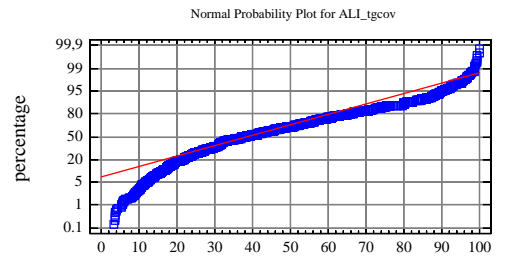
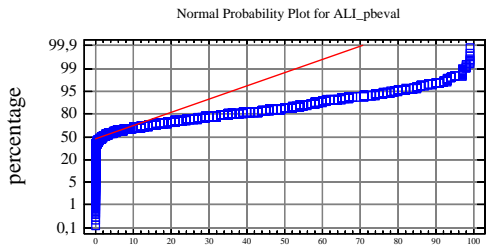
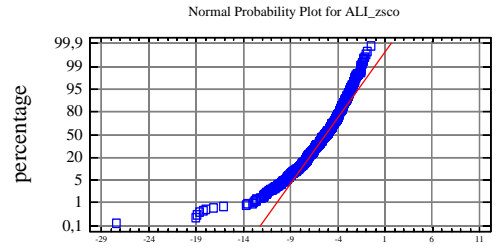
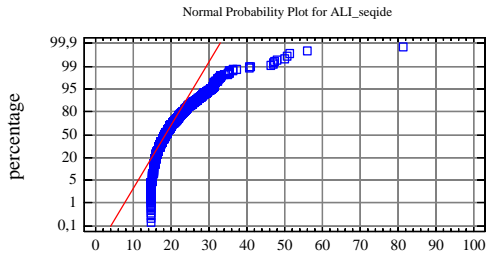


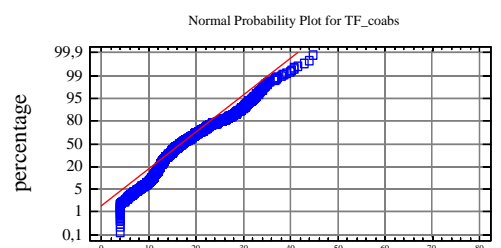
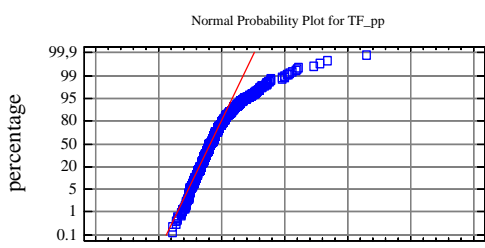
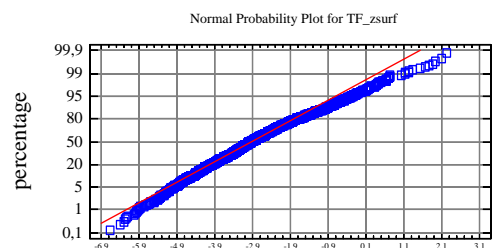
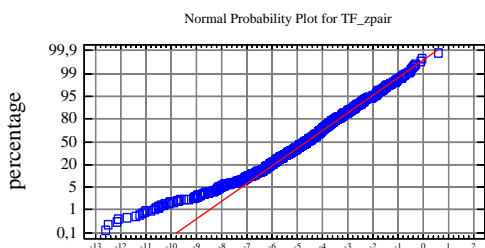
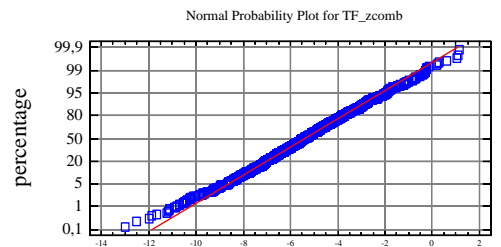
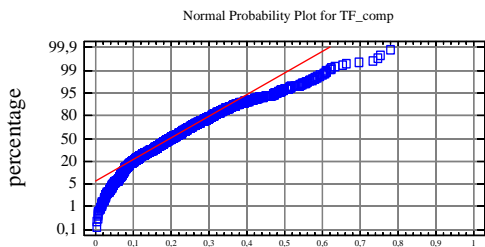
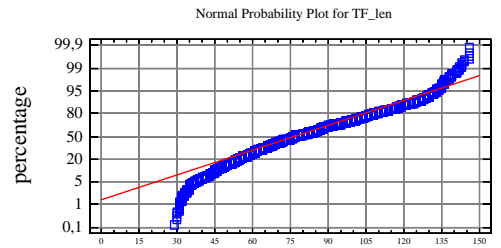
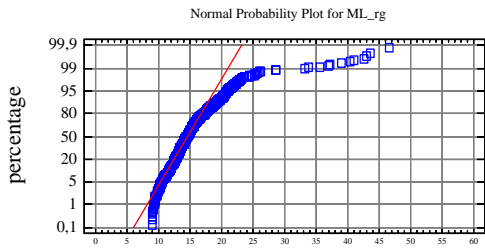
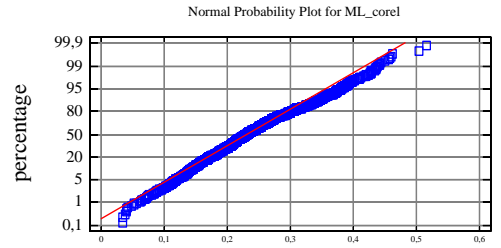
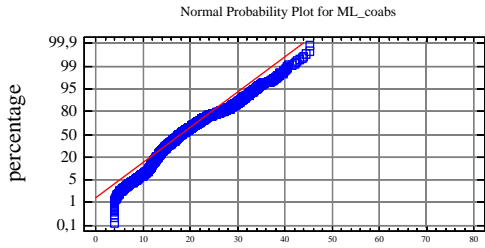


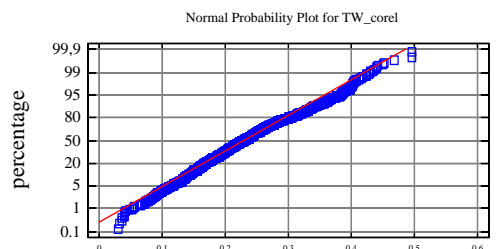
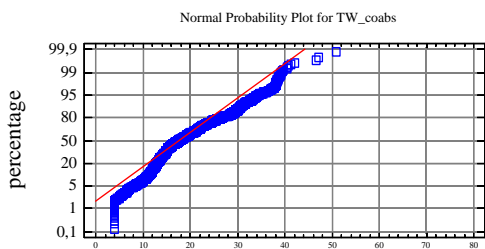
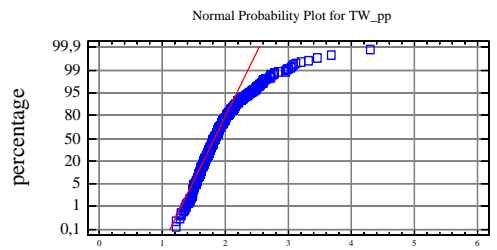
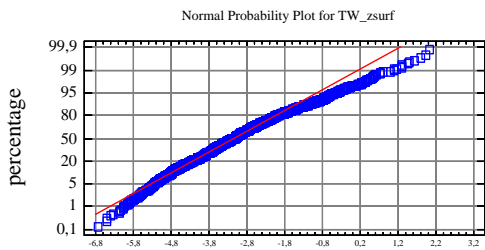
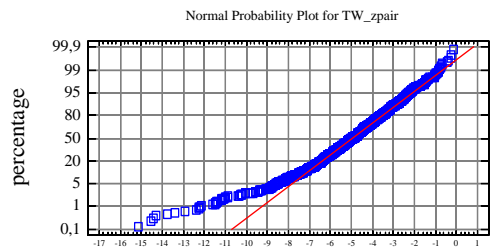
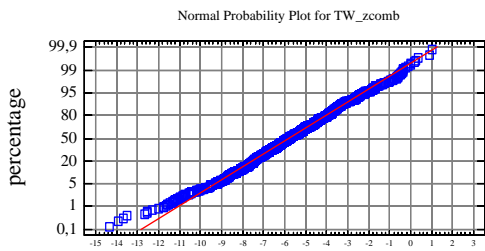
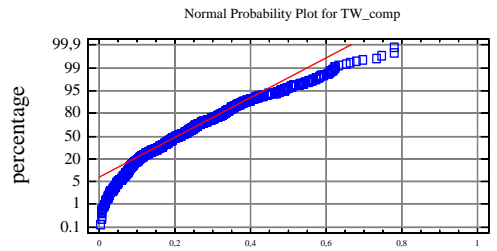
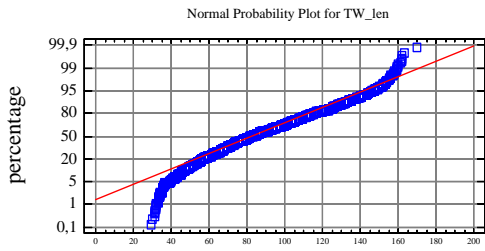
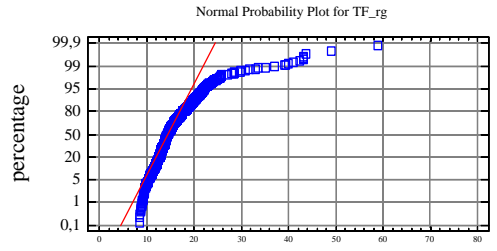
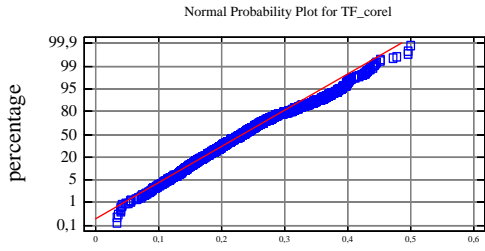


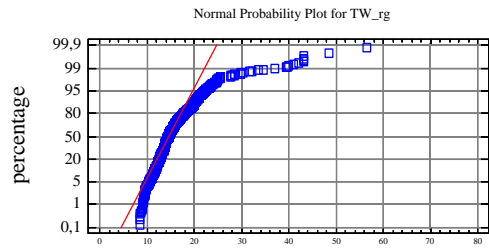


A continuación se presentan los gráficos de probabilidad normal (normal probability plots) sobre los modelos incorrectos:





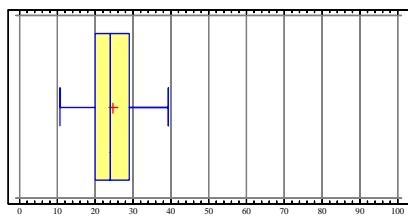




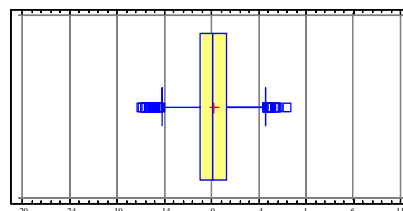
Anexo B: gráficos de caja y bigotes

A continuación se presentan los gráficos de caja y bigotes (box-and-whisker plots) sobre los modelos correctos:

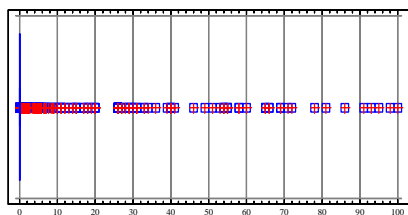
Box-and-Whisker Plot for ALL_seqide



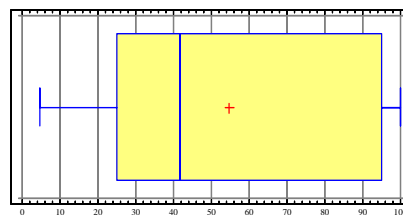
Box-and-Whisker Plot for ALL_zsco



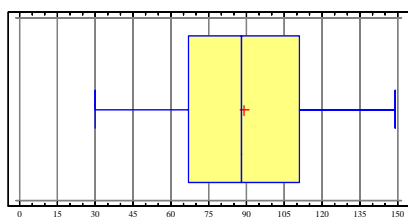
Box-and-Whisker Plot for ALL_pbeval



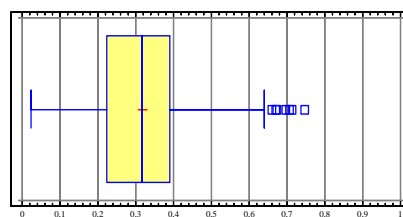
Box-and-Whisker Plot for ALL_tgcov



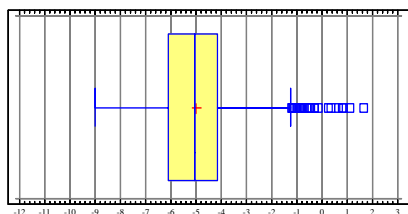
Box-and-Whisker Plot for ML_len



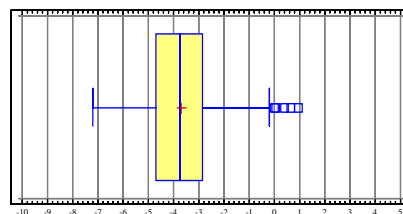
Box-and-Whisker Plot for ML_comp



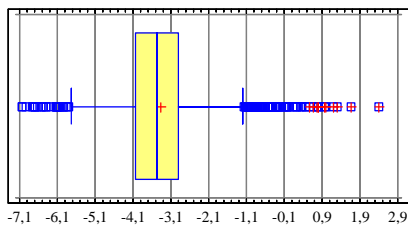
Box-and-Whisker Plot for ML_zcomb



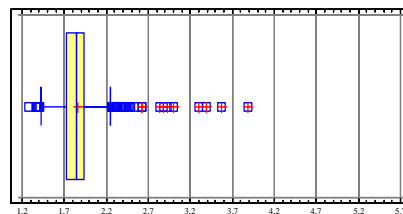
Box-and-Whisker Plot for ML_zpair

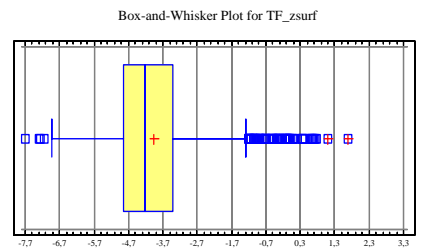
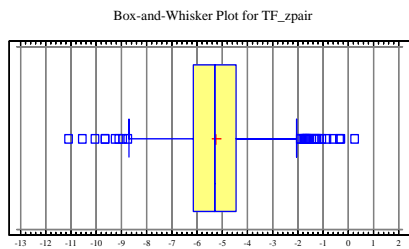
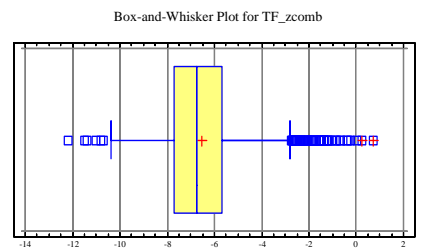
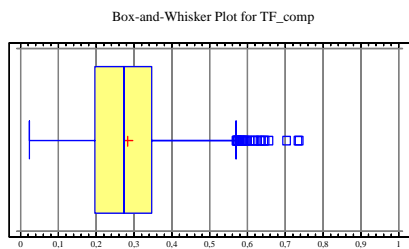
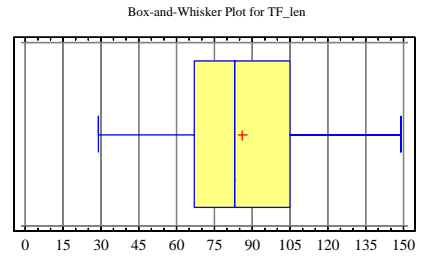
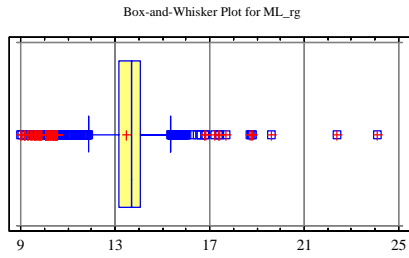
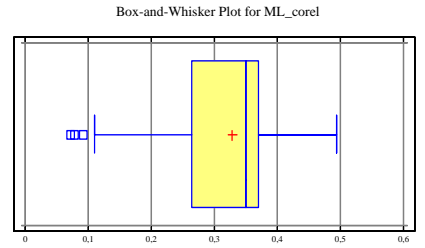
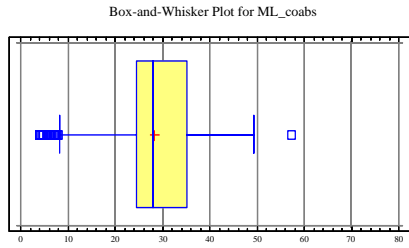


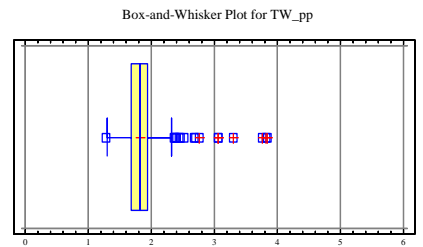
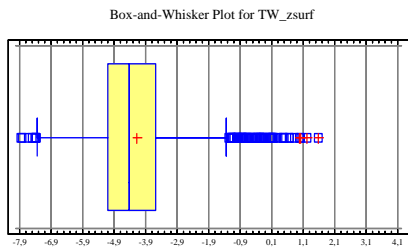
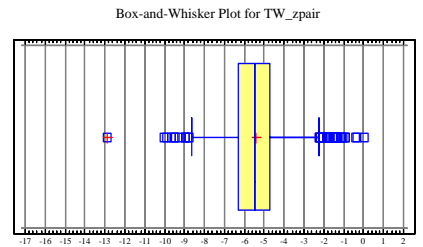
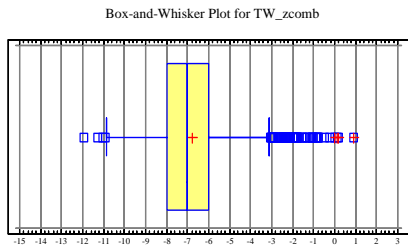
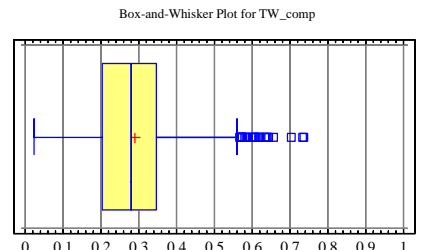
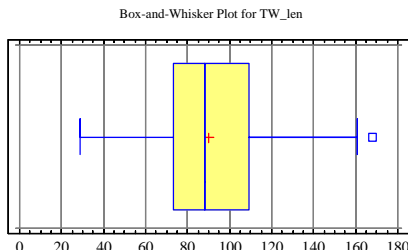
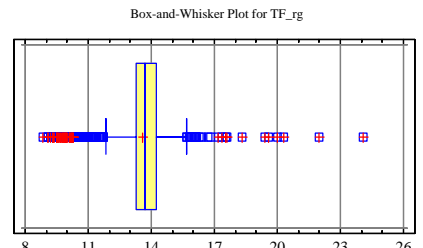
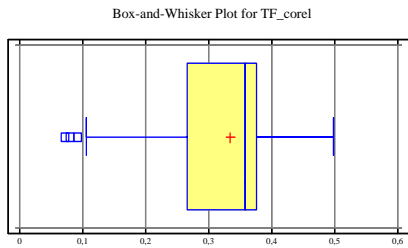
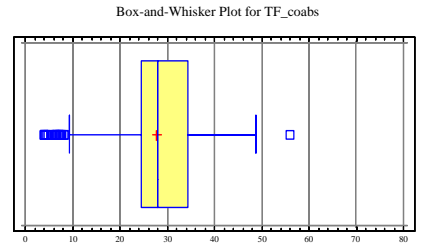
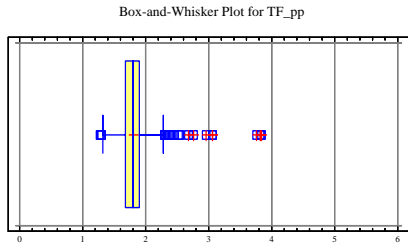
Box-and-Whisker Plot for ML_zsurf

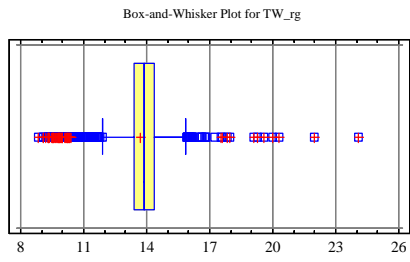
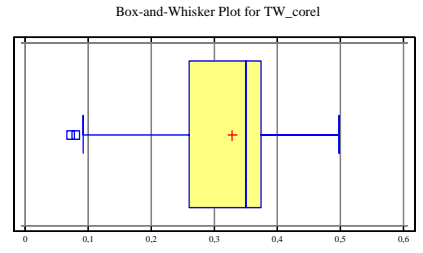
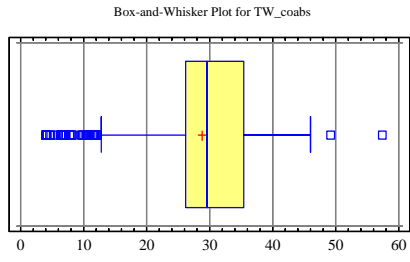


Box-and-Whisker Plot for ML_pp



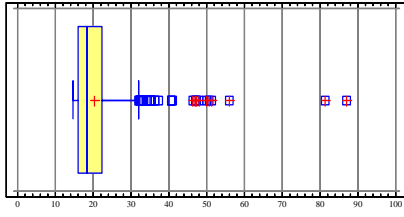




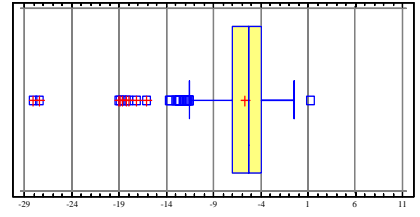


A continuación se presentan los gráficos de caja y bigotes (box-and-whisker plots) sobre los modelos incorrectos:

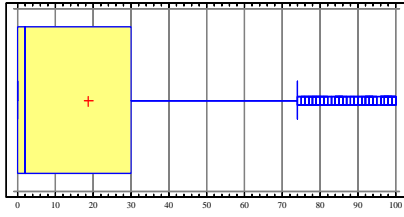
Box-and-Whisker Plot for ALL_seqide



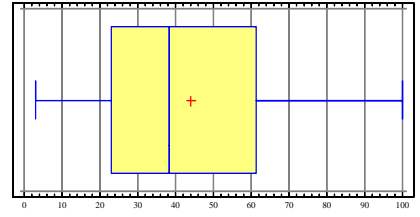
Box-and-Whisker Plot for ALL_zsco



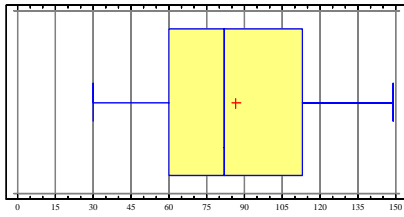
Box-and-Whisker Plot for ALL_pbeval



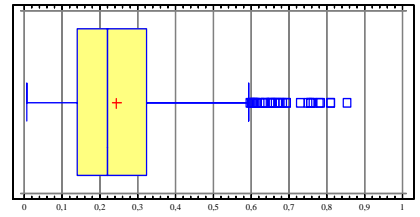
Box-and-Whisker Plot for ALL_tgcov



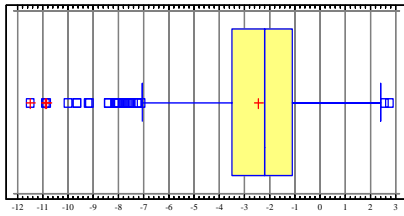
Box-and-Whisker Plot for ML_len



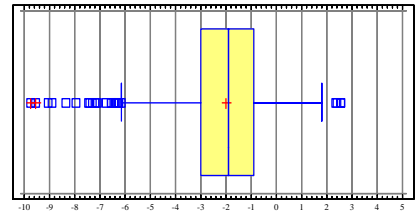
Box-and-Whisker Plot for ML_comp



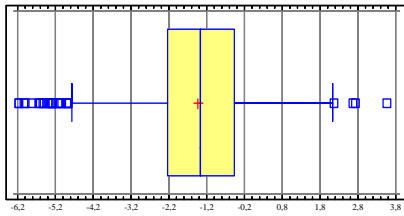
Box-and-Whisker Plot for ML_zcomb



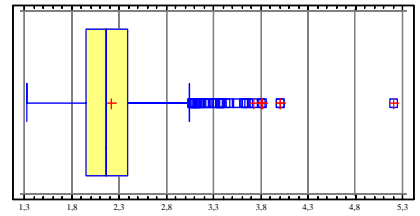
Box-and-Whisker Plot for ML_zpair

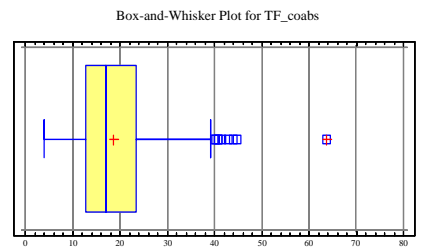
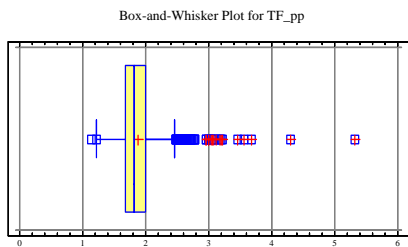
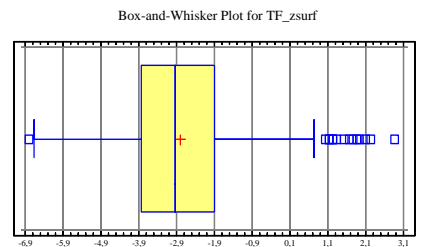
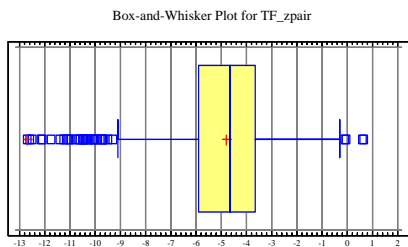
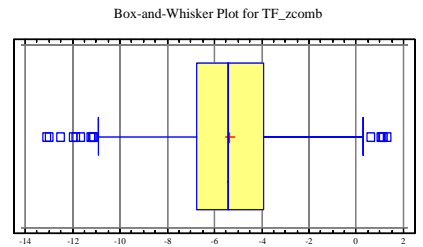
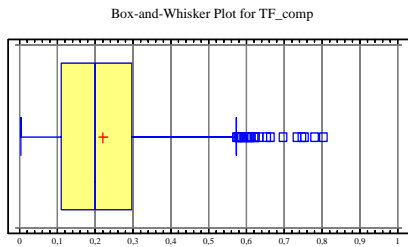
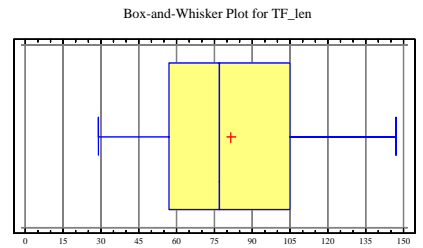
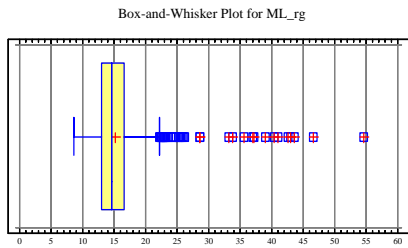
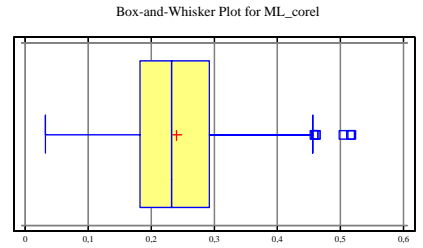
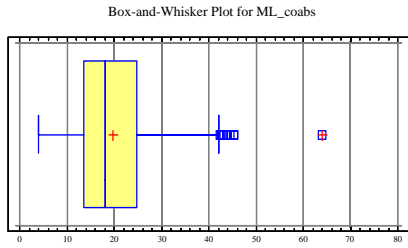


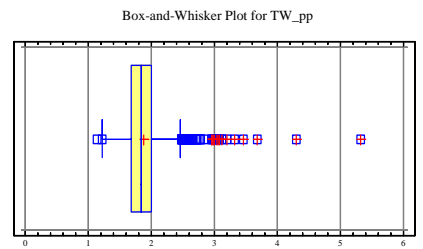
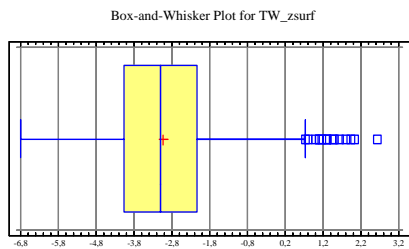
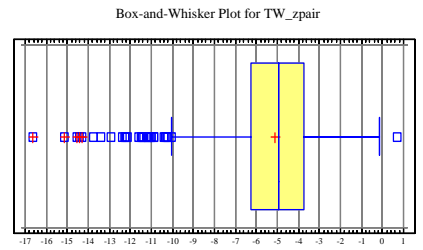
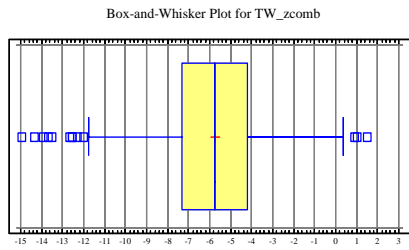
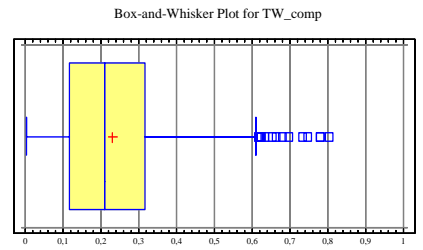
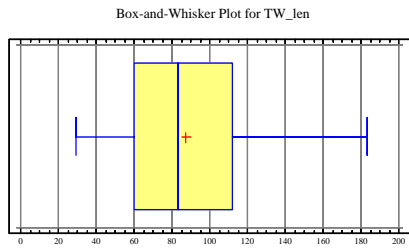
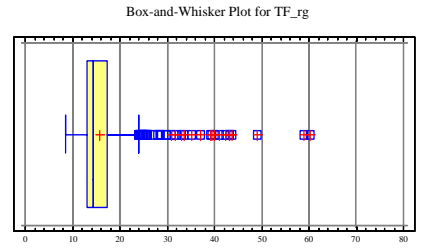
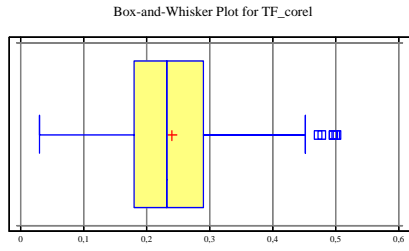
Box-and-Whisker Plot for ML_zsurf



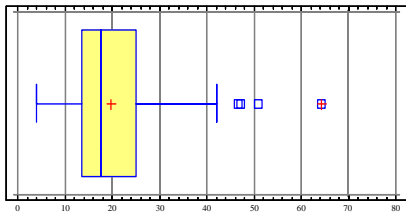
Box-and-Whisker Plot for ML_pp



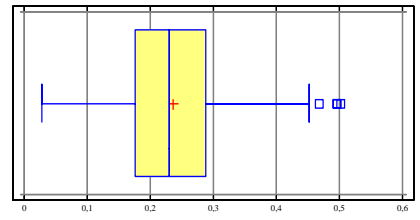




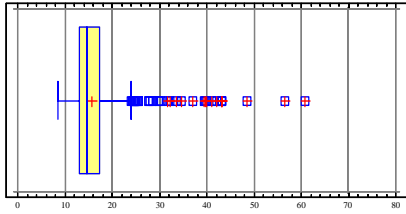
Box-and-Whisker Plot for TW_coabs



Box-and-Whisker Plot for TW_core1



Box-and-Whisker Plot for TW_rg



Anexo C: comportamiento de los índices de ranking medidos sobre las variables.

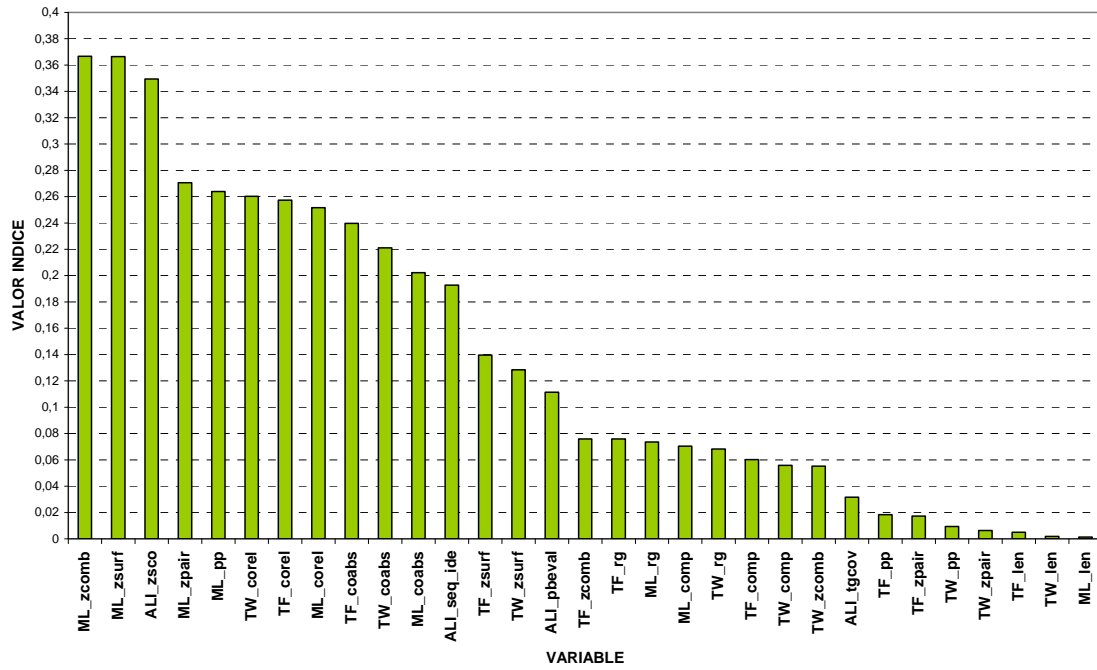


Figura C1. Valor del grado de relación funcional para cada variable.

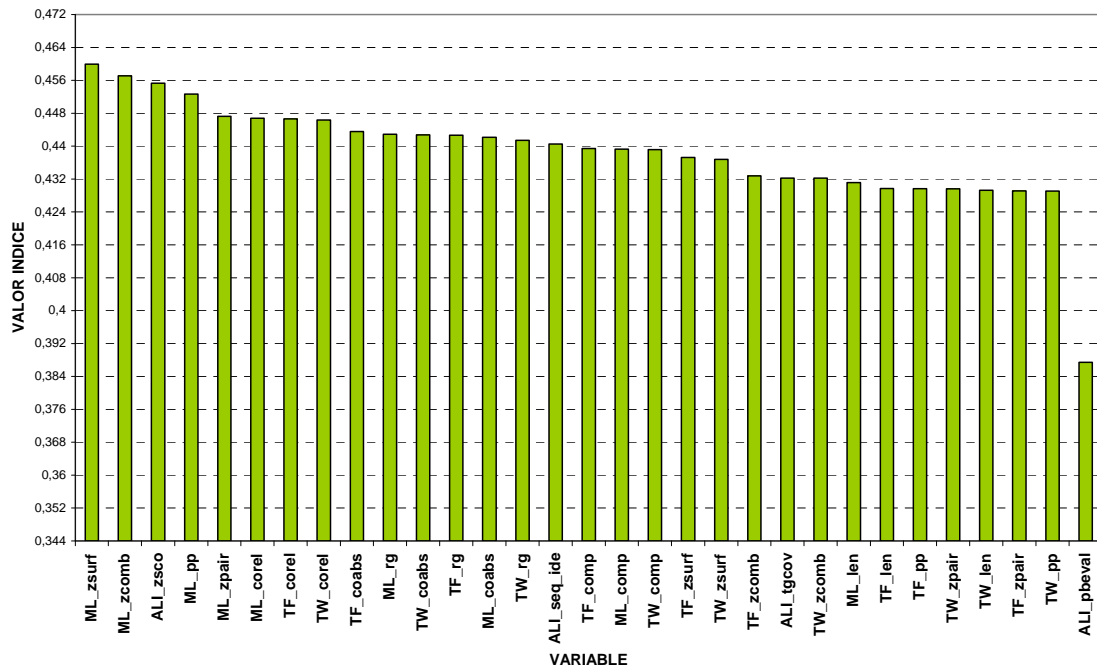


Figura C2. Valor del índice de razón de entropías para cada variable.

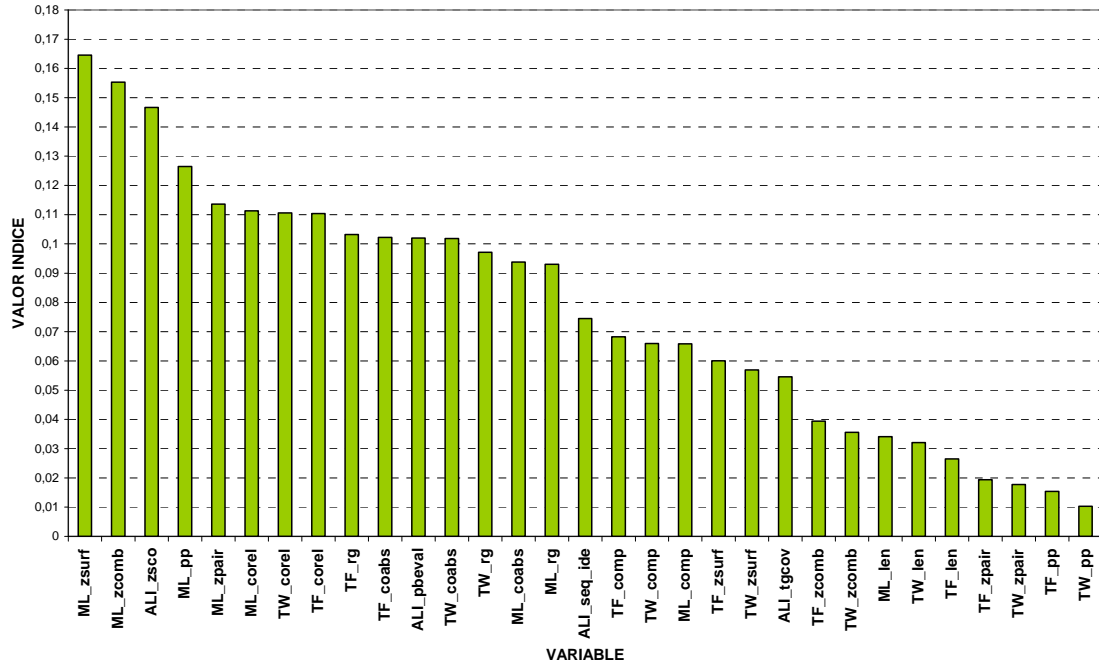


Figura C3. Valor del índice de incertidumbre simétrica para cada variable.

Anexo D: cantidad de información capturada por cada componente principal y tabla de pesos de las componentes principales.

Number of components extracted: 31

Principal Components Analysis

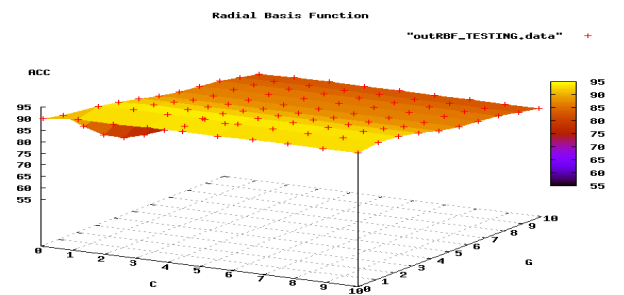
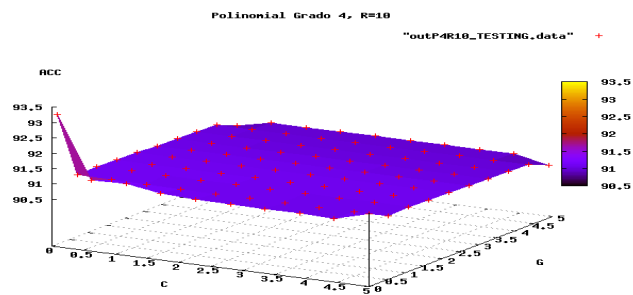
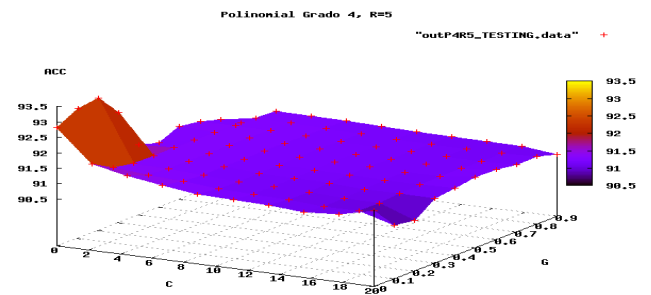
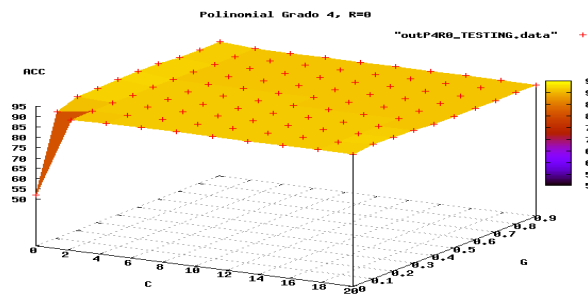
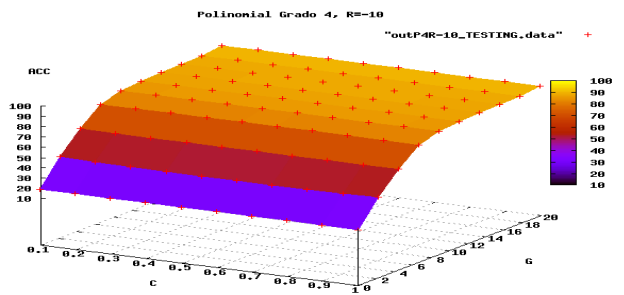
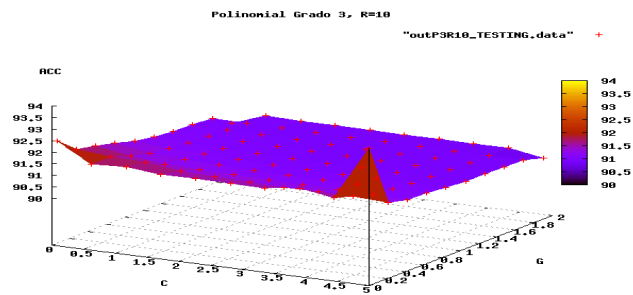
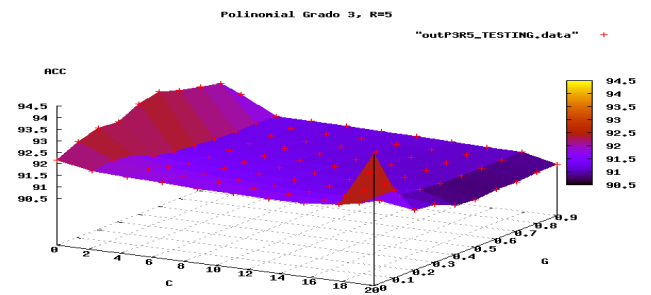
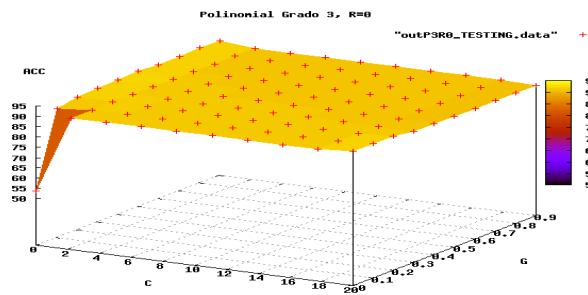
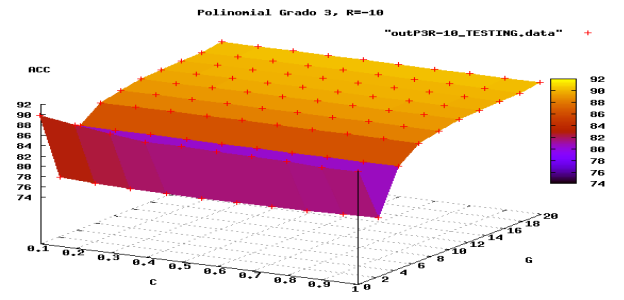
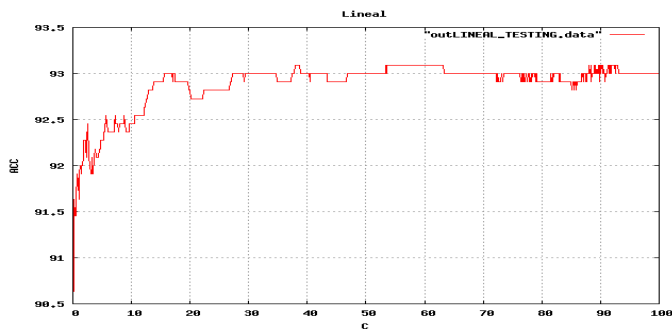
Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	11,612	37,458	37,458
2	6,70507	21,629	59,087
3	2,9331	9,462	68,549
4	1,8581	5,994	74,543
5	1,64167	5,296	79,838
6	1,29389	4,174	84,012
7	0,912351	2,943	86,955
8	0,77633	2,504	89,460
9	0,675576	2,179	91,639
10	0,608738	1,964	93,603
11	0,520378	1,679	95,281
12	0,454755	1,467	96,748
13	0,240735	0,777	97,525
14	0,186466	0,602	98,126
15	0,13709	0,442	98,568
16	0,114068	0,368	98,936
17	0,082399	0,266	99,202
18	0,0711342	0,229	99,432
19	0,0449827	0,145	99,577
20	0,034102	0,110	99,687
21	0,0255282	0,082	99,769
22	0,0203127	0,066	99,835
23	0,0110825	0,036	99,870
24	0,0095421	0,031	99,901
25	0,00836178	0,027	99,928
26	0,00793191	0,026	99,954
27	0,00631671	0,020	99,974
28	0,00407777	0,013	99,987
29	0,00203493	0,007	99,994
30	0,00149911	0,005	99,999
31	0,000406661	0,001	100,000

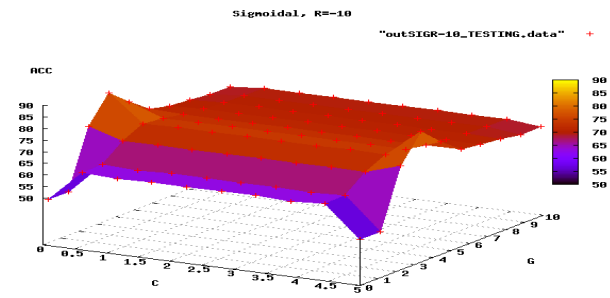
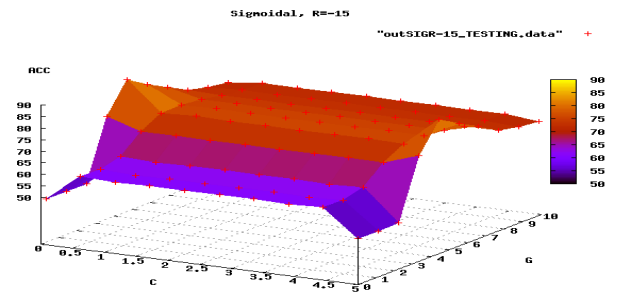
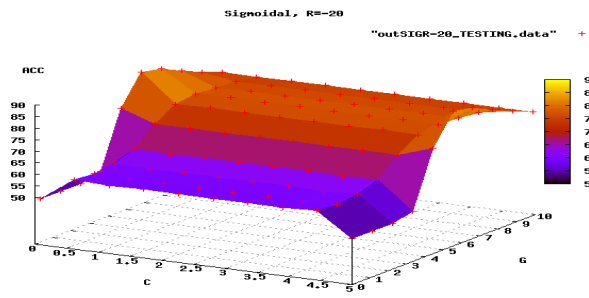
Variable	Componente Principal									
	1	2	3	4	5	6	7	8	9	10
ALI_seqide	0,080	0,106	-0,026	-0,462	-0,227	-0,139	-0,339	-0,467	-0,091	-0,057
ALI_zsco	-0,207	0,011	0,140	0,327	0,132	0,096	0,240	0,284	-0,023	0,138
ALI_pbeval	-0,141	0,004	0,022	0,076	-0,031	0,140	-0,694	0,414	0,222	0,071
ALI_tgcov	0,152	-0,071	0,102	-0,078	0,222	0,072	0,302	-0,262	0,390	0,565
ML_len	0,203	-0,238	0,034	0,001	0,061	0,246	-0,124	-0,034	0,062	-0,012
ML_comp	0,105	0,233	0,266	-0,130	0,220	0,222	-0,004	0,042	-0,229	-0,001
ML_zcomb	-0,247	-0,011	0,034	0,265	0,101	0,108	-0,129	-0,232	-0,083	-0,055
ML_zpair	-0,221	-0,003	-0,008	0,282	0,102	0,082	-0,089	-0,176	-0,014	-0,184
ML_zsurf	-0,236	-0,022	0,071	0,210	0,087	0,118	-0,147	-0,254	-0,138	0,086
ML_pp	-0,143	-0,099	0,034	0,204	0,308	-0,019	-0,020	-0,487	0,035	-0,274
ML_coabs	0,245	0,007	-0,235	0,085	0,127	0,203	-0,100	-0,055	0,035	0,062
ML_corel	0,083	0,282	-0,324	0,133	0,021	0,000	0,026	0,005	-0,160	0,073
ML_rg	-0,022	-0,343	-0,113	-0,070	-0,059	0,119	0,107	0,048	-0,300	0,004
TF_len	0,207	-0,237	0,035	-0,007	0,052	0,239	-0,112	-0,018	0,070	-0,028
TF_comp	0,098	0,237	0,299	-0,158	0,202	0,228	0,004	0,051	-0,248	-0,074
TF_zcomb	-0,256	0,056	-0,125	-0,161	-0,034	0,233	0,030	-0,038	0,067	0,070
TF_zpair	-0,207	0,081	-0,193	-0,204	-0,004	0,296	0,178	0,019	0,230	-0,272
TF_zsurf	-0,251	0,011	-0,024	-0,074	-0,059	0,111	-0,136	-0,087	-0,124	0,427
TF_pp	-0,111	-0,117	-0,164	-0,237	0,532	-0,248	-0,097	0,120	-0,018	-0,002
TF_coabs	0,248	0,016	-0,245	0,081	0,117	0,186	-0,092	-0,046	0,030	0,043
TF_corel	0,087	0,279	-0,329	0,133	0,020	-0,007	0,022	-0,004	-0,167	0,073
TF_rg	-0,036	-0,339	-0,127	-0,069	-0,071	0,112	0,096	0,051	-0,350	-0,007
TW_len	0,201	-0,242	0,033	0,020	0,041	0,228	-0,110	-0,026	0,072	-0,020
TW_comp	0,100	0,235	0,303	-0,144	0,202	0,229	0,003	0,046	-0,238	-0,067
TW_zcomb	-0,251	0,073	-0,124	-0,189	-0,024	0,230	0,053	-0,008	0,069	0,065
TW_zpair	-0,199	0,099	-0,193	-0,223	0,000	0,301	0,170	0,051	0,245	-0,253
TW_zsurf	-0,252	0,022	-0,021	-0,104	-0,039	0,087	-0,104	-0,081	-0,146	0,416
TW_pp	-0,104	-0,114	-0,181	-0,233	0,523	-0,277	-0,082	0,134	-0,031	0,006
TW_coabs	0,247	0,005	-0,245	0,096	0,109	0,181	-0,092	-0,049	0,033	0,027
TW_corel	0,086	0,281	-0,324	0,127	0,019	-0,016	0,021	-0,002	-0,170	0,058
TW_rg	-0,032	-0,340	-0,132	-0,068	-0,074	0,107	0,099	0,056	-0,345	-0,008

Variable	Componente Principal									
	11	12	13	14	15	16	17	18	19	20
ALI_seqide	-0,200	-0,057	-0,133	0,025	0,023	0,527	-0,051	0,047	0,074	0,004
ALI_zsco	0,025	0,086	-0,163	0,025	0,043	0,763	-0,088	0,058	0,104	0,013
ALI_pbeval	-0,409	0,274	-0,001	-0,009	0,017	-0,047	0,015	-0,027	0,017	-0,010
ALI_tgcov	-0,509	-0,001	-0,010	0,002	0,033	-0,062	0,067	-0,001	0,026	0,014
ML_len	0,150	-0,059	-0,018	0,146	0,013	0,118	0,258	0,048	-0,466	0,044
ML_comp	-0,088	0,000	0,095	-0,319	-0,656	0,037	-0,096	0,008	-0,204	-0,291
ML_zcomb	-0,117	-0,326	-0,098	0,022	-0,008	-0,061	0,013	-0,007	0,005	-0,008
ML_zpair	-0,254	-0,477	0,523	0,140	0,009	0,083	-0,021	0,030	0,004	-0,003
ML_zsurf	0,041	-0,121	-0,713	-0,091	-0,035	-0,204	0,023	-0,045	-0,004	0,002
ML_pp	0,036	0,698	0,168	-0,017	-0,005	0,013	0,023	0,025	0,003	-0,001
ML_coabs	0,078	-0,011	-0,027	0,119	0,010	0,050	-0,438	-0,053	-0,346	0,328
ML_corel	-0,105	0,056	0,002	-0,051	-0,029	0,045	0,227	0,032	-0,187	0,389
ML_rg	-0,198	0,079	-0,043	0,216	0,285	-0,008	-0,072	0,012	-0,242	-0,576
TF_len	0,158	-0,063	0,002	0,033	-0,040	0,114	0,326	0,051	-0,042	-0,022
TF_comp	-0,042	0,032	-0,002	0,169	0,294	-0,068	-0,001	-0,007	0,030	0,099
TF_zcomb	0,099	-0,003	0,101	-0,375	0,222	0,025	-0,021	-0,030	-0,113	-0,019
TF_zpair	-0,063	-0,035	-0,037	-0,333	0,170	0,073	0,045	-0,483	-0,097	-0,020
TF_zsurf	0,263	0,022	0,223	-0,290	0,198	-0,052	-0,095	0,501	-0,060	-0,005
TF_pp	0,041	-0,088	-0,055	-0,006	0,036	0,082	0,006	0,037	-0,125	-0,024
TF_coabs	0,086	-0,012	-0,015	0,052	0,013	0,025	-0,355	-0,052	0,153	-0,115
TF_corel	-0,100	0,060	-0,010	-0,020	0,042	0,032	0,261	0,037	0,010	-0,123
TF_rg	-0,230	0,073	0,018	-0,090	-0,125	-0,022	-0,038	-0,007	0,116	0,272
TW_len	0,164	-0,086	0,041	-0,200	-0,018	0,099	0,402	0,036	0,347	0,024
TW_comp	-0,043	0,019	0,028	0,103	0,350	-0,093	0,038	-0,007	0,189	0,157
TW_zcomb	0,107	0,053	0,005	0,369	-0,223	-0,025	0,079	0,038	0,126	0,012
TW_zpair	-0,055	0,013	-0,151	0,279	-0,190	-0,064	-0,018	0,475	0,082	0,022
TW_zsurf	0,258	0,080	0,179	0,349	-0,176	0,018	0,081	-0,508	0,067	0,004
TW_pp	0,050	-0,129	-0,016	-0,015	-0,010	-0,016	0,070	-0,011	0,120	0,034
TW_coabs	0,107	-0,034	0,014	-0,081	0,017	-0,024	-0,252	-0,044	0,456	-0,219
TW_corel	-0,094	0,058	-0,015	0,032	0,031	0,003	0,316	0,044	0,011	-0,256
TW_rg	-0,226	0,064	0,028	-0,111	-0,132	-0,035	-0,034	-0,001	0,153	0,261

Variable	Componente Principal										
	21	22	23	24	25	26	27	28	29	30	31
ALI_seqide	0,035	0,012	-0,017	0,007	0,006	0,002	0,009	0,003	0,008	-0,006	-0,001
ALI_zsco	0,031	0,046	-0,023	0,011	0,008	0,009	0,013	0,002	0,011	-0,008	-0,001
ALI_pbeval	0,013	0,003	0,000	0,005	0,002	-0,003	0,000	0,001	0,001	-0,001	0,000
ALI_tgcov	-0,012	0,015	0,007	0,012	0,005	0,001	0,002	-0,001	0,003	-0,008	-0,001
ML_len	-0,065	0,331	-0,342	0,003	-0,039	-0,034	-0,154	-0,014	0,109	-0,380	0,230
ML_comp	0,072	-0,061	-0,028	-0,026	-0,004	-0,003	0,010	-0,001	-0,004	0,008	0,003
ML_zcomb	0,005	0,002	0,020	0,004	-0,003	0,057	0,002	-0,792	0,015	0,020	-0,012
ML_zpair	-0,016	-0,008	0,007	0,001	-0,001	-0,032	-0,002	0,431	-0,011	-0,009	0,005
ML_zsurf	0,004	-0,009	0,004	-0,009	-0,003	-0,034	0,008	0,425	-0,004	-0,010	0,006
ML_pp	0,024	0,007	0,003	-0,001	0,005	-0,001	0,001	-0,001	0,003	0,002	0,000
ML_coabs	0,126	-0,068	-0,176	0,112	-0,015	0,000	-0,037	0,003	-0,112	0,453	-0,270
ML_corel	0,251	-0,451	0,180	0,072	0,065	0,062	0,152	-0,004	0,093	-0,340	0,202
ML_rg	0,266	-0,317	0,020	-0,004	0,008	0,003	0,007	0,002	-0,016	0,014	0,011
TF_len	-0,056	0,054	0,481	-0,118	0,083	0,072	0,444	0,021	0,075	0,050	-0,444
TF_comp	-0,039	0,155	0,401	0,445	0,097	-0,041	-0,326	0,003	-0,133	-0,064	-0,011
TF_zcomb	0,026	0,079	0,111	0,065	-0,645	0,395	-0,010	0,035	-0,002	-0,016	-0,005
TF_zpair	-0,020	0,027	-0,047	-0,034	0,394	-0,260	0,038	-0,017	0,006	0,007	-0,001
TF_zsurf	-0,007	0,033	-0,027	-0,029	0,357	-0,223	0,036	-0,016	-0,001	0,005	0,003
TF_pp	-0,605	-0,334	0,000	0,046	-0,027	0,004	-0,038	-0,007	0,047	0,004	0,003
TF_coabs	-0,106	0,139	0,393	-0,372	-0,055	-0,110	0,023	-0,016	-0,008	0,009	0,542
TF_corel	-0,078	0,110	0,009	-0,413	-0,135	-0,204	-0,470	-0,009	-0,082	-0,040	-0,419
TF_rg	-0,184	0,139	-0,078	-0,049	0,002	0,039	0,166	-0,017	-0,646	-0,175	-0,008
TW_len	0,128	-0,359	-0,109	0,077	-0,030	-0,026	-0,315	0,003	-0,187	0,329	0,225
TW_comp	-0,012	-0,107	-0,382	-0,418	-0,090	0,043	0,307	0,001	0,139	0,058	0,007
TW_zcomb	0,020	-0,067	-0,031	0,195	-0,388	-0,597	0,192	-0,045	-0,001	-0,008	-0,004
TW_zpair	0,009	-0,019	-0,016	-0,128	0,217	0,388	-0,155	0,026	-0,007	0,013	0,009
TW_zsurf	-0,028	-0,040	-0,005	-0,104	0,196	0,329	-0,128	0,022	0,005	0,006	0,001
TW_pp	0,601	0,324	-0,003	-0,049	0,023	-0,007	0,036	0,007	-0,047	-0,005	-0,003
TW_coabs	-0,018	-0,056	-0,244	0,308	0,056	0,104	0,048	0,002	0,122	-0,458	-0,268
TW_corel	-0,177	0,319	-0,169	0,320	0,086	0,151	0,304	0,023	-0,013	0,382	0,209
TW_rg	-0,065	0,148	0,049	0,059	-0,010	-0,040	-0,165	0,014	0,667	0,164	-0,002

Anexo E: gráficos en el espacio de parámetros de SVM





Anexo F: valores de $\mathbf{Y}_i \alpha_i^*$ con su vector de soporte asociado \mathbf{x}_i .

Cada vector de soporte \mathbf{x}_i corresponde a un vector de la forma (1:...22:...).

$$\mathbf{b}^* = -0.418318$$

$\mathbf{Y}_i \alpha_i^*$	\mathbf{x}_i
0.06516514565233249	1:-0.357567 2:0.157044 3:-0.636364 4:-0.565695 5:0.378151 6:-0.579039
7:0.383349 8:0.178487 9:0.636504 10:-0.0992366 11:-0.464602 12:-0.517908 13:-0.544403 14:0.128205	
15:-0.64053 16:0.484651 17:0.454005 18:0.465829 19:-0.466667 20:-0.5 21:-0.451648 22:-0.612513	
0.08054799406216374	1:-0.632296 2:0.486143 3:-0.909091 4:-0.565695 5:0.378151 6:-0.960263
7:0.304003 8:0.33465 9:0.23585 10:-0.394402 11:-0.986726 12:-0.982421 13:0.418075 14:0.367521	
15:-0.96189 16:0.518278 17:0.337767 18:0.655322 19:-0.187302 20:-1 21:-0.984062 22:0.220962	
1.06643827891807	1:-0.672842 2:0.393764 3:-1 4:-0.751737 5:-0.310924 6:-0.104681 7:0.243302
8:0.340113 9:0.117418 10:-0.496183 11:-0.429204 12:-0.101296 13:-0.813174 14:-0.299145 15:-	
0.143041 16:-0.118404 17:0.109152 18:-0.273116 19:-0.726984 20:-0.419643 21:-0.0954124 22:-	
0.84519	
0.1704673969219876	1:-0.270185 2:0.0450346 3:0.878788 4:0.054651 5:-0.159664 6:-0.867627
7:0.475987 8:0.549898 9:0.319012 10:-0.236641 11:-0.725664 12:-0.586904 13:-0.421533 14:-0.128205	
15:-0.844728 16:0.0924746 17:0.166883 18:-0.0316674 19:-0.504762 20:-0.723214 21:-0.595089 22:-	
0.568106	
0.03593284475512563	1:0.441454 2:0.187067 3:0.111111 4:-0.830551 5:-0.915966 6:-0.934434
7:0.29618 8:0.293955 9:0.260845 10:-0.628499 11:-0.973451 12:-0.575478 13:-0.410008 14:-0.897436	
15:-0.899575 16:0.567272 17:0.56991 18:0.438279 19:-0.847619 20:-0.883929 21:-0.330174 22:-	
0.512707	
0.2215948691876383	1:0.241524 2:-0.2806 3:-0.993333 4:-0.408068 5:0.260504 6:-0.411648 7:0.630356
8:0.577814 9:0.63374 10:-0.506361 11:-0.278761 12:-0.301252 13:-0.70957 14:0.230769 15:-0.425261	
16:0.0464585 17:0.332216 18:-0.206638 19:-0.466667 20:-0.200893 21:-0.210424 22:-0.775494	
1.463631676771209	1:-0.590353 2:0.180139 3:-1 4:-0.0737322 5:0.327731 6:0.366447 7:-0.201423 8:-
0.156912 9:-0.160612 10:-0.796438 11:0.0929204 12:0.00637223 13:-0.770153 14:0.350427 15:0.302948	
16:-0.372532 17:0.0259395 18:-0.672154 19:-0.739683 20:0.129464 21:0.0251992 22:-0.822777	
3.07899727701259	1:-0.0506816 2:-0.644342 3:-1 4:0.582495 5:0.563025 6:-0.529616 7:-0.172485 8:-
0.252935 9:0.0374807 10:-0.62341 11:0.49115 12:0.19666 13:-0.609526 14:0.452991 15:-0.680958 16:-	
0.181079 17:0.0442079 18:-0.325715 19:-0.587302 20:0.446429 21:0.234547 22:-0.594922	
0.4968341355967436	1:-0.621811 2:0.655889 3:0.474747 4:-0.941305 5:-0.882353 6:-0.875574
7:0.576523 8:0.585244 9:0.5022 10:-0.659033 11:-0.995575 12:-0.660294 13:-0.623822 14:-0.863248	
15:-0.854513 16:0.409516 17:0.583165 18:0.132454 19:-0.180952 20:-0.995536 21:-0.663149 22:-	
0.703346	
0.4868231124531356	1:-0.738553 2:0.682448 3:0.919192 4:-0.876594 5:-0.579832 6:-0.88377
7:0.377526 8:0.271957 9:0.479647 10:-0.389313 11:-0.969027 12:-0.774116 13:-0.368623 14:-0.555556	
15:-0.876915 16:0.561471 17:0.624067 18:0.343211 19:-0.015873 20:-0.96875 21:-0.771699 22:-	
0.534398	
0.3726721115937138	1:-0.717581 2:0.600462 3:-1 4:-0.711708 5:0.193277 6:-0.997765 7:0.130091
8:0.180308 9:0.134972 10:-0.592875 11:-0.995575 12:-0.973193 13:0.805693 14:0.230769 15:-0.991245	
16:0.729733 17:0.75522 18:0.560197 19:-0.0031746 20:-1 21:-0.972001 22:0.37255	
0.4436207848727248	1:-0.396714 2:-0.0889145 3:-1 4:0.76107 5:0.94958 6:-0.258661 7:-0.413846 8:-
0.274136 9:-0.502338 10:-0.648855 11:-0.163717 12:-0.45902 13:-0.607325 14:0.760684 15:-0.301661	
16:-0.45181 17:-0.33946 18:-0.426862 19:-0.739683 20:-0.214286 21:-0.437002 22:-0.711992	
0.3013773309851215	1:0.197483 2:0.174365 3:-0.999996 4:0.170383 5:0.159664 6:-0.74643 7:0.156354
8:0.214548 9:0.103896 10:-0.715013 11:0.0265487 12:0.0586684 13:-0.531285 14:-0.333333	
15:0.402601 16:-0.0976578 17:0.304868 18:-0.450208 19:-0.492063 20:-0.383929 21:-0.0122765 22:-	
0.894431	
1.354770728890706	1:0.464523 2:-0.542725 3:-1 4:0.798818 5:1 6:-0.35999 7:-0.599212 8:-0.620281
9:-0.482158 10:-0.62341 11:-0.367257 12:-0.608438 13:-0.591607 14:0.965812 15:-0.377108 16:-	
0.833634 17:-0.673368 18:-0.672268 19:-0.530159 20:-0.330357 21:-0.57269 22:-0.676199	
3.253736864071019	1:-0.0485844 2:0.140878 3:-1 4:0.032044 5:-0.0252101 6:0.391034 7:-0.257286 8:-
0.0902026 9:-0.297012 10:-0.679389 11:-0.261062 12:-0.117996 13:-0.803636 14:-0.025641	
15:0.296768 16:-0.228542 17:0.0993768 18:-0.49423 19:-0.669841 20:-0.28125 21:-0.133319 22:-	
0.852776	
0.5256214745708058	1:-0.34778 2:0.0288684 3:-1 4:0.673131 5:0.831933 6:0.171737 7:-0.187954
8:0.024648 9:-0.415167 10:-0.648855 11:-0.221239 12:-0.466051 13:-0.660246 14:0.760684 15:-	
0.185786 16:-0.298971 17:-0.0969885 18:-0.475504 19:-0.492063 20:-0.236607 21:-0.455094 22:-	
0.699003	
2.811071557161565	1:-0.312129 2:0.136259 3:-1 4:0.838847 5:0.092437 6:0.458587 7:-0.184807 8:-
0.0228758 9:-0.244811 10:-0.659033 11:0.106195 12:0.192705 13:-0.822226 14:0.042735 15:0.401313	
16:-0.442343 17:-0.179536 18:-0.441801 19:-0.549206 20:0.0714286 21:0.198363 22:-0.858391	
0.279205655621145	1:0.366655 2:-0.19515 3:-0.995354 4:-0.258944 5:0.378151 6:-0.577052 7:0.277564
8:0.219783 9:0.389804 10:-0.51145 11:0.707965 12:0.520545 13:-0.583199 14:0.196581 15:-0.667825	
16:-0.0138568 17:0.224424 18:-0.193668 19:-0.644444 20:0.732143 21:0.718716 22:-0.668968	

0.6434391056321167 1:-0.714785 2:0.48037 3:-0.999859 4:-0.994815 5:-0.714286 6:0.517447
7:0.145193 8:0.137076 9:0.195789 10:-0.6743 11:-0.362832 12:0.568446 13:-0.927375 14:-0.692308
15:0.472898 16:0.173579 17:0.469932 18:-0.156614 19:-0.777778 20:-0.40625 21:0.440448 22:-
0.939353
0.7863020624671061 1:-0.471513 2:0.284065 3:-0.999939 4:-0.991496 5:-0.680672 6:0.933441
7:0.164794 8:0.190014 9:0.132645 10:-0.56743 11:-0.256637 12:0.714348 13:-0.947276 14:-0.65812
15:0.429896 16:0.0681884 17:0.329606 18:-0.117818 19:-0.52381 20:-0.263393 21:0.655395 22:-
0.923792
0.8929251224908508 1:-0.166725 2:0.325635 3:-0.999999 4:-1 5:-0.764706 6:0.354526 7:0.279622
8:0.266933 9:0.313852 10:-0.628499 11:-0.411504 12:0.574599 13:-0.954352 14:-0.74359 15:0.259946
16:0.0430312 17:0.375775 18:-0.348001 19:-0.784127 20:-0.397321 21:0.569244 22:-0.951585
0.9008709912615691 1:-0.0842363 2:0.32679 3:-0.999999 4:-0.996474 5:-0.731092 6:0.666956
7:0.139444 8:0.119956 9:0.124582 10:-0.704835 11:-0.415929 12:0.497253 13:-0.936956 14:-0.709402
15:0.464143 16:0.0149805 17:0.276272 18:-0.264463 19:-0.784127 20:-0.388393 21:0.517123 22:-
0.950588
0.01322985733872855 1:-0.357567 2:0.247113 3:-0.59596 4:-0.868713 5:-0.12605 6:-0.80231
7:0.246758 8:0.263259 9:0.214794 10:-0.709924 11:-0.283186 12:-0.0687761 13:-0.531589 14:-
0.213675 15:-0.7837 16:-0.256874 17:0.101955 18:-0.490689 19:-0.479365 20:-0.40625 21:-0.146672
22:-0.634001
0.4502750182430132 1:-0.105907 2:0.279446 3:-0.999879 4:-0.841128 5:0.0588235 6:-0.847759
7:0.153766 8:0.418431 9:-0.227487 10:-0.913486 11:-0.584071 12:-0.518348 13:-0.400704 14:0.025641
15:-0.841895 16:-0.153956 17:0.151999 18:-0.505893 19:-0.542857 20:-0.558036 21:-0.469309 22:-
0.533202
0.1685318578024151 1:-0.707095 2:0.469977 3:-0.59596 4:-0.78347 5:0.445378 6:-0.911834 7:0.275711
8:0.293256 9:0.273123 10:-0.587786 11:-0.309735 12:-0.405845 13:-0.125897 14:0.213675 15:-0.91245
16:0.989746 17:1 18:0.770385 19:-0.161905 20:-0.464286 21:-0.453371 22:-0.343768
0.0538227599676625 1:-0.616218 2:0.418014 3:0.191919 4:-0.803588 5:0.310924 6:-0.870856
7:0.401832 8:0.357364 9:0.465595 10:-0.577608 11:-0.19469 12:-0.246319 13:-0.292939 14:0.0769231
15:-0.952363 16:0.339046 17:0.605988 18:-0.0983158 19:-0.307937 20:-0.75 21:-0.690717 22:-
0.154349
1.591354229845239 1:0.216358 2:0.181293 3:-0.890909 4:-0.732448 5:-0.798319 6:-0.344344
7:0.471091 8:0.447485 9:0.429588 10:0.0636132 11:-0.747788 12:-0.137332 13:-0.872856 14:-0.82906
15:-0.391013 16:0.515736 17:0.532219 18:0.429626 19:0.231746 20:-0.732143 21:-0.0424295 22:-
0.897223
0.04810629013589305 1:-0.340091 2:0.19515 3:0.79798 4:-0.462201 5:-0.12605 6:-0.525394 7:0.482443
8:0.432608 9:0.502569 10:-0.440204 11:-0.951327 12:-0.873874 13:-0.504686 14:-0.179487 15:-
0.422943 16:0.450687 17:0.3356 18:0.523938 19:-0.530159 20:-0.986607 21:-0.899634 22:-0.663062
1.453039664643845 1:0.206571 2:-0.337182 3:-0.999998 4:-0.529192 5:0.411765 6:-0.103191 7:-
0.088921 8:0.17795 9:-0.299846 10:-0.465649 11:-0.0442478 12:-0.16414 13:-0.711925 14:0.384615
15:-0.766705 16:-0.151596 17:0.130235 18:-0.278513 19:-0.428571 20:-0.1875 21:-0.274176 22:-
0.537609
2.173701836663117 1:-0.702202 2:0.203233 3:-1 4:-0.725189 5:-0.277311 6:-0.528871 7:0.199321
8:0.312409 9:0.0392545 10:-0.659033 11:-0.278761 12:0.0700945 13:-0.764079 14:-0.247863 15:-
0.481396 16:0.401805 17:0.718556 18:0.0198146 19:-0.580952 20:-0.303571 21:0.0144303 22:-0.790442
1.14854241815186 1:-0.686823 2:0.30254 3:-1 4:-0.667738 5:-0.0756303 6:-0.425059 7:0.140091
8:0.108802 9:0.166601 10:-0.24173 11:0.0840708 12:0.325862 13:-0.775684 14:-0.0940171 15:-
0.478306 16:0.0396741 17:0.3809 18:-0.310473 19:-0.453968 20:0.191964 21:0.469309 22:-0.815288
0.002137163031070107 1:-0.65187 2:0.32679 3:-1 4:-0.620035 5:0.092437 6:-0.326214 7:0.000485251
8:-0.0438819 9:0.103435 10:-0.40458 11:0.367257 12:0.464733 13:-0.741476 14:0.128205 15:-0.357281
16:-0.046585 17:0.238296 18:-0.353284 19:-0.6 20:0.388393 21:0.446909 22:-0.794214
0.1825810581254686 1:-0.63719 2:0.133949 3:0.535354 4:-0.428808 5:0.764706 6:-0.614802 7:0.206541
8:0.397815 9:-0.0365362 10:-0.399491 11:0.615044 12:0.167655 13:-0.531067 14:0.538462 15:-
0.808678 16:-0.280135 17:0.0786885 18:-0.434795 19:-0.447619 20:0.339286 21:0.0876588 22:-
0.564002
0.2641477142450142 1:-0.00803915 2:-0.367206 3:-0.999798 4:-0.457638 5:0.663866 6:-0.755371
7:0.00135282 8:0.109404 9:-0.172153 10:-0.521628 11:0.327434 12:0.00856955 13:-0.475945
14:0.333333 15:-0.907815 16:-0.305727 17:-0.33636 18:-0.165419 19:-0.498413 20:0.0892857 21:-
0.00107689 22:-0.445332
2.104052344775787 1:-0.730164 2:0.247113 3:-1 4:-0.715649 5:-0.243697 6:-0.412393 7:0.158986
8:0.320635 9:-0.0436085 10:-0.526718 11:-0.30531 12:0.00373544 13:-0.752847 14:-0.213675 15:-
0.381486 16:0.166471 17:0.414036 18:-0.140482 19:-0.625397 20:-0.303571 21:-0.0144303 22:-0.79652
2.689198068120338 1:-0.0730514 2:-0.45612 3:-1 4:-0.50057 5:0.512605 6:-0.704955 7:-0.0212334 8:-
0.08526 9:0.0829091 10:-0.526718 11:0.269912 12:0.0468029 13:-0.546434 14:0.350427 15:-0.74662
16:-0.0545352 17:0.212609 18:-0.302767 19:-0.530159 20:0.0.191964 21:0.0833513 22:-0.647564
0.2015043473792784 1:-0.632995 2:0.0311778 3:-1 4:-0.419268 5:0.798319 6:-0.824662 7:-0.157913
8:-0.129727 9:-0.112811 10:-0.475827 11:0.371681 12:-0.027027 13:-0.376386 14:0.82906 15:-
0.855285 16:-0.205478 17:-0.00412818 18:-0.362372 19:-0.657143 20:0.375 21:-0.041568 22:-0.468902
0.5554012819994504 1:-0.72597 2:0.505774 3:-1 4:0.810225 5:0.663866 6:-0.510245 7:1 8:0.872435
9:1 10:-0.83715 11:-0.168142 12:-0.371567 13:-0.517368 14:0.675214 15:-0.667053 16:0.583088
17:0.546359 18:0.535582 19:-0.790476 20:-0.477679 21:-0.608874 22:-0.603624
0.1461093296936231 1:-0.215659 2:0.297921 3:-0.232323 4:-0.280929 5:-0.411765 6:-0.0438346
7:0.281975 8:0.411082 9:0.0877929 10:-0.0839695 11:-0.579646 12:-0.237091 13:-0.813056 14:-
0.487179 15:-0.198146 16:0.173073 17:0.4693 18:-0.0719026 19:0.0349206 20:-0.683036 21:-0.329313
22:-0.840959
0.6067473072335844 1:0.137365 2:0.147806 3:-0.850505 4:-0.488334 5:-0.697479 6:-0.547995
7:0.493971 8:0.285029 9:0.752678 10:-0.277354 11:-0.20354 12:0.862448 13:-0.804281 14:-0.675214

15:-0.611948 16:0.294926 17:0.450446 18:0.0822028 19:-0.657143 20:-0.165179 21:0.892742 22:-0.824388
0.1586579433394009 1:0.294652 2:-0.0508083 3:0.979798 4:-0.594525 5:-0.210084 6:-0.495343
7:0.330838 8:0.412285 9:0.238983 10:-0.582697 11:-0.699115 12:-0.535926 13:-0.727399 14:-0.230769
15:-0.469036 16:0.310208 17:0.433443 18:0.16131 19:-0.68254 20:-0.745536 21:-0.577428 22:-0.764184
3.6 1:-0.507165 2:-0.0461894 3:-0.99996 4:0.981748 5:0.394958 6:-0.460822 7:-0.0425992
8:0.0321757 9:-0.177475 10:-0.689567 11:0.681416 12:0.482312 13:-0.646616 14:0.0598291 15:-0.310673
16:-0.0440847 17:0.160762 18:-0.247953 19:-0.530159 20:0.352679 21:0.480078 22:-0.786003
3.6 1:-0.370849 2:0.0658199 3:-0.999596 4:0.854402 5:0.277311 6:-0.240531 7:-0.19513 8:-0.297938
9:0.0469027 10:-0.659033 11:0.482301 12:0.404526 13:-0.731299 14:0.179487 15:-0.0848461 16:-0.0632721
17:0.187445 18:-0.255375 19:-0.574603 20:0.473214 21:0.485247 22:-0.784647
3.6 1:-0.51835 2:-0.0138568 3:-0.982626 4:0.981748 5:0.394958 6:-0.115112 7:-0.020557 8:-0.00175593
9:0.00428483 10:-0.679389 11:0.5 12:0.326302 13:-0.755559 14:0.299145 15:-0.369383
16:-0.0662078 17:0.280669 18:-0.4841 19:-0.612698 20:0.607143 21:0.496446 22:-0.777325
1.868426284408159 1:-0.686823 2:0.407621 3:-0.232323 4:0.472363 5:-0.0756303 6:-0.377127
7:0.046393 8:0.0813091 9:-0.0108733 10:-0.735369 11:0.132743 12:0.382114 13:-0.736404 14:-0.0598291
15:-0.509978 16:-0.00299891 17:0.295489 18:-0.263441 19:-0.669841 20:0.160714
21:0.395649 22:-0.782693
0.2905952228912683 1:0.241524 2:-0.150115 3:-0.945455 4:-0.0137924 5:0.0252101 6:-0.562399 7:-0.126327
8:-0.195509 9:0.0375268 10:-0.613232 11:-0.269912 12:-0.159306 13:-0.584174 14:0.0598291
15:-0.928158 16:0.105678 17:0.348032 18:-0.116474 19:-0.631746 20:-0.285714 21:-0.197932 22:-0.346005
3.502874492734526 1:-0.386229 2:0.0254042 3:-1 4:0.0561029 5:0.12605 6:-0.372408 7:-0.2362 8:-0.269258
9:-0.0980672 10:-0.526718 11:0.128319 12:0.18875 13:-0.746282 14:0.128205 15:-0.325093
16:-0.108614 17:0.309598 18:-0.532458 19:-0.695238 20:0.142857 21:0.200948 22:-0.797269
3.6 1:0.29605 2:-0.28291 3:-0.993737 4:-0.00197034 5:0.0420168 6:-0.543773 7:0.126886 8:0.155741
9:0.0800756 10:-0.547074 11:-0.292035 12:-0.196221 13:-0.646019 14:-0.025641 15:-0.63641
16:0.225185 17:0.35863 18:0.0999441 19:-0.752381 20:-0.370536 21:-0.237993 22:-0.682763
3.6 1:-0.0744495 2:0.0473441 3:-0.999939 4:-0.153583 5:-0.176471 6:-0.112877 7:-0.116769 8:-0.100202
9:-0.0837154 10:-0.435115 11:0.0707965 12:0.417271 13:-0.762688 14:-0.264957 15:-0.284408
16:0.116522 17:0.396764 18:-0.12738 19:-0.644444 20:0.00446429 21:0.441309 22:-0.816273
0.8461678719330791 1:-0.217756 2:-0.0508083 3:-1 4:0.102769 5:0.193277 6:-0.265864 7:-0.000323501
8:-0.00370696 9:0.0651708 10:-0.424936 11:0.30531 12:0.306966 13:-0.739158 14:0.162393 15:-0.280803
16:0.0716297 17:0.38623 18:-0.303733 19:-0.587302 20:0.325893 21:0.354728 22:-0.783215
3.6 1:-0.309332 2:0.0531178 3:-0.999998 4:-0.060251 5:-0.0420168 6:-0.499317 7:-0.148237 8:-0.135727
9:-0.0736253 10:-0.506361 11:0.265487 12:0.496374 13:-0.712 14:-0.0769231 15:-0.421656
16:-0.169421 17:0.173921 18:-0.443429 19:-0.593651 20:0.272321 21:0.542537 22:-0.772942
3.6 1:-0.372947 2:0.144342 3:-0.999996 4:-0.0137924 5:0.0252101 6:-0.313548 7:-0.175029 8:-0.128053
9:-0.196134 10:-0.40458 11:0.132743 12:0.280598 13:-0.724388 14:-0.0598291 15:-0.203811
16:-0.0906486 17:0.14534 18:-0.298469 19:-0.631746 20:0.0491071 21:0.268576 22:-0.798916
3.6 1:0.00943726 2:-0.0357968 3:-1 4:0.114383 5:0.210084 6:-0.0900286 7:-0.253904 8:-0.353982 9:-0.0657237
10:-0.618321 11:0.455752 12:0.438805 13:-0.733388 14:0.179487 15:-0.140466 16:-0.0250799
17:0.238501 18:-0.376876 19:-0.52381 20:0.40625 21:0.418479 22:-0.794785
0.1271799725699767 1:-0.270884 2:-0.017321 3:0.717172 4:-0.382765 5:0.529412 6:-0.551968
7:0.423007 8:0.347088 9:0.499205 10:-0.430025 11:-0.340708 12:-0.462096 13:-0.555923 14:0.435897
15:-0.472898 16:-0.237532 17:0.0627293 18:-0.463197 19:-0.619048 20:-0.321429 21:-0.410295 22:-0.6652
0.9082246270254853 1:0.0716533 2:0.0496536 3:-0.941414 4:-0.30727 5:-0.0252101 6:-0.332174
7:0.47537 8:0.577407 9:0.303992 10:-0.445293 11:0.0309735 12:0.21336 13:-0.782669 14:0.00854701
15:-0.453843 16:-0.0833725 17:0.131153 18:-0.219551 19:-0.6 20:-0.0625 21:0.0799052 22:-0.763139
1.194356555272193 1:-0.507165 2:0.637413 3:-0.935354 4:-0.830965 5:-0.563025 6:0.0473116
7:0.269829 8:0.0959906 9:0.554033 10:-0.557252 11:-0.384956 12:0.258185 13:-0.901197 14:-0.538462
15:0.185271 16:0.297398 17:0.429726 18:0.158072 19:-0.625397 20:-0.397321 21:0.206978 22:-0.912509
0.6519211899643478 1:-0.444949 2:0.459584 3:-0.434343 4:-0.796329 5:-0.428571 6:-0.0488017
7:0.250596 8:0.0328748 9:0.55401 10:-0.557252 11:-0.384956 12:0.0775654 13:-0.869329 14:-0.401709
15:-0.340286 16:0.418281 17:0.404087 18:0.386419 19:-0.663492 20:-0.401786 21:0.02563 22:-0.866251
0.2815805971222856 1:-0.509962 2:0.450346 3:-0.999091 4:-0.639946 5:0.613445 6:-0.979635
7:0.587654 8:0.670422 9:0.406321 10:-0.302799 11:-0.814159 12:-0.865085 13:0.291112 14:0.230769
15:-0.991245 16:0.766815 17:0.785382 18:0.593805 19:-0.0031746 20:-1 21:-0.972001 22:0.37255
0.7393990278291278 1:-0.1108 2:0.108545 3:-0.945455 4:-0.556984 5:-0.0756303 6:-0.112132
7:0.138238 8:0.3081 9:-0.077864 10:-0.608142 11:-0.623894 12:-0.505164 13:-0.74608 14:-0.0598291
15:-0.276426 16:0.0434105 17:0.109183 18:-0.0730765 19:-0.663492 20:-0.620536 21:-0.500754 22:-0.779478
0.714411472554108 1:-0.590353 2:0.51963 3:-0.737374 4:-0.342113 5:-1 6:-0.118341 7:0.428903
8:0.362387 9:0.375429 10:-0.628499 11:-0.597345 12:0.763568 13:-0.974872 14:-0.982906 15:-0.204068
16:0.372631 17:0.599978 18:0.0638366 19:-0.498413 20:-0.584821 21:0.764376 22:-0.970743
0.3066899443210362 1:-0.403006 2:0.554273 3:-0.414141 4:-0.31764 5:-0.983193 6:-0.214951
7:0.343146 8:0.397262 9:0.254302 10:1 11:-0.60177 12:0.698528 13:-0.911326 14:-0.965812 15:-0.643105
16:0.435685 17:0.495967 18:0.291691 19:-0.663492 20:-0.566964 21:0.766961 22:-0.878758
0.4525742282171453 1:-0.455435 2:0.513857 3:-0.999996 4:-0.0247848 5:-0.781513 6:-0.414628
7:0.0995206 8:0.0826911 9:-0.0061508 10:-0.628499 11:-0.59292 12:0.195342 13:-0.889848 14:-0.760684
15:-0.356251 16:0.161021 17:0.367487 18:-0.0489165 19:-0.403175 20:-0.607143 21:0.143657
22:-0.904102

1.274113799649271 1:0.416987 2:-0.13164 3:-0.737374 4:-0.559058 5:-0.260504 6:-0.404446
7:0.482634 8:0.453826 9:0.494321 10:-0.287532 11:-0.256637 12:0.0872336 13:-0.739387 14:-0.264957
15:-0.491953 16:0.302117 17:0.496663 18:0.11246 19:-0.587302 20:-0.254464 21:0.0967047 22:-
0.737775
0.1531534265564264 1:-0.724572 2:1 3:0.69697 4:-0.0932282 5:-0.478992 6:-0.91705 7:0.567038
8:0.567912 9:0.51259 10:-0.572519 11:-0.880531 12:-0.667765 13:-0.421602 14:-0.452991 15:-
0.908845 16:0.42633 17:0.410714 18:-0.454373 19:-0.765079 20:-0.897321 21:-0.695456 22:-0.548022
2.499334128881552 1:-0.00803915 2:0.140878 3:-0.737374 4:-0.744063 5:-0.378151 6:-0.451385
7:0.364659 8:0.379995 9:0.336382 10:-0.643766 11:-0.376106 12:0.0366952 13:-0.752735 14:-0.401709
15:-0.446891 16:0.233248 17:0.410872 18:0.0166526 19:-0.390476 20:-0.401786 21:0.0264915 22:-
0.810629
1.350433007713754 1:0.16323 2:0.214781 3:-0.717172 4:-0.51737 5:-0.647059 6:-0.676642 7:0.222333
8:0.0621403 9:0.398719 10:-0.750636 11:-0.871681 12:-0.553065 13:-0.658269 14:-0.623932 15:-
0.73941 16:0.396945 17:0.592417 18:0.105037 19:-0.555556 20:-0.870536 21:-0.558044 22:-0.72398
0.875454173449548 1:-0.307934 2:-0.054273 3:-1 4:-0.135539 5:0.89916 6:-0.0932572 7:-0.0407611
8:0.079667 9:-0.218319 10:-0.577608 11:-0.0221239 12:-0.347836 13:-0.709506 14:0.794872 15:-
0.152826 16:-0.455968 17:-0.361382 18:-0.277566 19:-0.333333 20:-0.0535714 21:-0.331467 22:-
0.774386
3.6 1:-0.477805 2:0.285219 3:-0.963636 4:-0.463652 5:0.0588235 6:-0.620017 7:-0.0456577
8:0.0589211 9:-0.141814 10:-0.475827 11:0.190265 12:0.310481 13:-0.698589 14:0.042735 15:-
0.522081 16:-0.160417 17:-0.0426579 18:-0.23112 19:-0.619048 20:0.09375 21:0.219901 22:-0.756962
0.2266502576888546 1:-0.0856344 2:0.0542725 3:1 4:-0.396661 5:0.394958 6:-0.0957407 7:0.135326
8:0.0095763 9:0.303601 10:-0.776081 11:-0.132743 12:-0.233575 13:-0.721367 14:0.401709 15:-
0.106734 16:0.0380728 17:-0.120998 18:0.240796 19:-0.68254 20:-0.178571 21:-0.275038 22:-0.770286
0.08056866330135468 1:-0.382733 2:0.233256 3:0.292929 4:-0.648657 5:-0.563025 6:-0.572582
7:0.0879775 8:0.0122264 9:0.181115 10:-0.582697 11:-0.681416 12:-0.268732 13:-0.7891 14:-0.538462
15:-0.517446 16:0.300755 17:0.544556 18:0.0447699 19:-0.625397 20:-0.683036 21:-0.291837 22:-
0.822466
1.033209095783453 1:-0.607829 2:0.211316 3:-0.757576 4:-0.0212589 5:-0.378151 6:0.000124177
7:0.0806841 8:0.0832927 9:0.172499 10:-0.552163 11:-0.429204 12:-0.0415293 13:-0.787134 14:-
0.350427 15:-0.511523 16:0.26363 17:0.407171 18:0.143853 19:-0.11746 20:-0.410714 21:-0.0376911
22:-0.763259
2.578977426036366 1:-0.16323 2:-0.0542725 3:-0.998222 4:0.167272 5:-0.176471 6:-0.305849
7:0.00424963 8:-0.0850161 9:0.157663 10:-0.592875 11:-0.460177 12:-0.247198 13:-0.78488 14:-
0.145299 15:-0.464658 16:-0.246353 17:-0.0672688 18:-0.430119 19:-0.44127 20:-0.504464 21:-
0.31639 22:-0.802135
2.246691472083665 1:-0.396714 2:0.495381 3:-0.965657 4:-0.508452 5:-0.89916 6:-0.174469 7:0.27202
8:0.296882 9:0.244396 10:-0.653944 11:-0.526549 12:0.640079 13:-0.947895 14:-0.880342 15:-
0.441741 16:0.0100221 17:0.239988 18:-0.205256 19:-0.606349 20:-0.526786 21:0.601551 22:-0.932294
1.873482361678721 1:-0.396714 2:0.393764 3:-0.939394 4:-0.508452 5:-0.89916 6:0.017509 7:0.288048
8:0.263503 9:0.321869 10:-0.475827 11:-0.570796 12:0.522742 13:-0.966293 14:-0.880342 15:0.104159
16:0.260919 17:0.51471 18:0.0482917 19:-0.644444 20:-0.553571 21:0.528322 22:-0.958951
1.358578892947355 1:0.279273 2:0.0127021 3:-0.999982 4:0.277196 5:-0.0588235 6:-0.551223 7:-
0.143487 8:-0.171414 9:-0.0553341 10:-0.557252 11:-0.482301 12:-0.345199 13:-0.711099 14:-
0.145299 15:-0.532381 16:-0.0294764 17:0.20081 18:-0.185281 19:-0.790476 20:-0.504464 21:-
0.321559 22:-0.766705
0.7823887012400418 1:0.182104 2:-0.249423 3:-0.717172 4:0.847765 5:0.193277 6:-0.630945 7:-
0.033747 8:-0.0288915 9:0.115183 10:-0.618321 11:-0.123894 12:-0.11448 13:-0.667205 14:0.162393
15:-0.650315 16:-0.174464 17:-0.110227 18:-0.112952 19:-0.606349 20:-0.142857 21:-0.112643 22:-
0.703091
0.7158886219067614 1:0.362461 2:0.128176 3:0.959596 4:-0.852536 5:-0.94958 6:-0.742705 7:0.294827
8:0.169479 9:0.441637 10:-0.267176 11:-0.814159 12:-0.0507581 13:-0.785083 14:-0.931624 15:-
0.72087 16:0.340169 17:0.473665 18:0.0983347 19:0.466667 20:-0.799107 21:-0.029076 22:-0.819331
0.3806127701204991 1:-0.628102 2:0.45843 3:0.616162 4:-0.765841 5:-0.94958 6:-0.282255 7:0.559451
8:0.52125 9:0.550877 10:-0.832061 11:-0.90708 12:-0.337728 13:-0.860303 14:-0.965812 15:-0.539848
16:0.663504 17:0.492503 18:0.748781 19:-0.155556 20:-0.9375 21:-0.386173 22:-0.846825
0.01870276960981837 1:-0.492485 2:0.513857 3:-0.971717 4:-0.242559 5:-0.00840336 6:-0.921768
7:0.639311 8:0.555734 9:0.683361 10:0.221374 11:-0.712389 12:-0.632169 13:-0.176431 14:-0.0769231
15:-0.928415 16:0.339916 17:0.633241 18:-0.00201649 19:-0.27619 20:-0.700893 21:-0.585613 22:-
0.324334
0.1727877725760315 1:-0.492485 2:0.513857 3:-0.971717 4:-0.242559 5:-0.00840336 6:-0.924252
7:0.537232 8:0.464329 9:0.589187 10:0.287532 11:-0.707965 12:-0.626895 13:-0.13528 14:-0.0769231
15:-0.9279 16:0.23204 17:0.504777 18:-0.11441 19:-0.326984 20:-0.727679 21:-0.617489 22:-0.322403
0.189124904935848 1:0.0178259 2:-0.0207852 3:-0.656566 4:-0.685782 5:-0.378151 6:-0.585744
7:0.435417 8:0.252967 9:0.64095 10:-0.475827 11:-0.774336 12:-0.555702 13:-0.670466 14:-0.384615
15:-0.769023 16:0.157285 17:0.326 18:0.0190383 19:-0.593651 20:-0.84375 21:-0.645488 22:-0.643269
0.2388692426009162 1:-0.142258 2:0.206697 3:0.717172 4:-0.715026 5:-0.445378 6:-0.363964 7:0.1615
8:0.188112 9:0.175471 10:-0.704835 11:-0.570796 12:-0.193144 13:-0.803263 14:-0.606838
15:0.413158 16:0.0423149 17:0.272049 18:-0.101743 19:-0.498413 20:-0.611607 21:-0.0893819 22:-
0.896461
0.4388045083291058 1:-0.567284 2:0.0300231 3:-1 4:-0.596184 5:0.226891 6:-0.0110518 7:-0.0210423
8:-0.118054 9:0.150476 10:-0.547074 11:-0.168142 12:-0.176884 13:-0.750939 14:0.179487 15:-
0.142526 16:-0.202936 17:-0.428303 18:0.155421 19:-0.352381 20:-0.125 21:-0.104028 22:-0.805345
0.3594891927937011 1:0.00594198 2:-0.438799 3:0.777778 4:-0.820595 5:0.613445 6:-0.68608
7:0.326648 8:0.53172 9:0.00882306 10:-0.709924 11:0.331858 12:0.0402109 13:-0.546231 14:0.487179
15:-0.74971 16:-0.157355 17:0.0792262 18:-0.40007 19:-0.269841 20:0.285714 21:-0.0743054 22:-
0.583835

0.632314330382949 1:-0.729465 2:0.383372 3:-0.985253 4:-0.578762 5:0.462185 6:-0.397988 7:-0.00164691 8:-0.0607746 9:0.0937133 10:-0.689567 11:0.221239 12:0.0380136 13:-0.678735 14:0.367521 15:-0.400541 16:0.223443 17:0.386088 18:0.0609776 19:-0.536508 20:-0.138393 21:-0.221193 22:-0.758481
3.6 1:-0.0367005 2:0.274827 3:-0.757576 4:-0.684331 5:-0.411765 6:-0.0704085 7:0.167956 8:0.0579943 9:0.300468 10:-0.531807 11:-0.225664 12:0.306087 13:-0.841451 14:-0.452991 15:0.105704 16:0.429713 17:0.214428 18:0.0106694 19:-0.707937 20:-0.196429 21:0.41331 22:-0.880098
0.04718695607892946 1:-0.691716 2:0.678984 3:-0.951515 4:-0.851084 5:-0.731092 6:-0.922762 7:0.311208 8:0.267828 9:0.298486 10:-0.572519 11:-1 12:-0.767084 13:-0.427229 14:-0.709402 15:-0.929188 16:0.794445 17:0.928397 18:0.476772 19:-0.0793651 20:-1 21:-0.767823 22:-0.528332
2.403314372515181 1:0.22964 2:0.165127 3:-0.999899 4:-0.666286 5:-0.142857 6:-0.137464 7:0.047437 8:-0.00170715 9:0.0918012 10:-0.618321 11:-0.539823 12:-0.369809 13:-0.759086 14:-0.111111 15:-0.100811 16:0.323117 17:0.50397 18:0.0326331 19:-0.593651 20:-0.540179 21:-0.384019 22:-0.795072
0.06442336142077669 1:0.222649 2:0.130485 3:-0.0707071 4:-0.714819 5:-0.613445 6:0.0311685 7:0.517631 8:0.429714 9:0.574996 10:0.221374 11:-0.637168 12:-0.133817 13:-0.85003 14:-0.606838 15:-0.292648 16:0.600098 17:0.723238 18:0.314715 19:-0.542857 20:-0.611607 21:-0.0936894 22:-0.857996
0.07825024323706017 1:-0.51835 2:0.191686 3:-1 4:-0.294618 5:0.428571 6:-0.268099 7:0.0788019 8:0.0965272 9:0.110392 10:-0.577608 11:-0.39823 12:-0.47528 13:-0.709373 14:0.350427 15:-0.0554912 16:-0.176795 17:0.0172719 18:-0.269897 19:-0.809524 20:-0.397321 21:-0.443463 22:-0.775738
0.5764585725227593 1:-0.616218 2:0.199769 3:-1 4:-0.158975 5:0.764706 6:0.472991 7:-0.173646 8:-0.23453 9:0.0531917 10:-0.552163 11:0.221239 12:-0.120193 13:-0.762843 14:0.777778 15:0.551178 16:-0.39009 17:-0.141987 18:-0.483115 19:-0.771429 20:0.169643 21:-0.165626 22:-0.802805
3.6 1:0.324712 2:-0.202079 3:-1 4:0.119154 5:-0.092437 6:0.444431 7:-0.264609 8:-0.250415 9:-0.265521 10:-0.648855 11:-0.185841 12:0.0208745 13:-0.804947 14:-0.0598291 15:0.207158 16:-0.0609264 17:0.217702 18:-0.420784 19:-0.536508 20:-0.183036 21:0.000646134 22:-0.847168
0.6537238772308137 1:0.0870325 2:-0.236721 3:-1 4:0.488541 5:0.344538 6:-0.390786 7:-0.415493 8:-0.4502 9:-0.205418 10:-0.62341 11:-0.486726 12:-0.526258 13:-0.648769 14:0.350427 15:-0.356508 16:-0.0805352 17:0.0909465 18:-0.119958 19:-0.422222 20:-0.446429 21:-0.487831 22:-0.706317
1.699552835664348 1:-0.0863334 2:-0.0635104 3:-1 4:0.829514 5:0.747899 6:-0.021731 7:-0.369732 8:-0.26573 9:-0.407842 10:-0.516539 11:-0.274336 12:-0.481433 13:-0.670887 14:0.760684 15:-0.0364362 16:-0.377912 17:-0.106985 18:-0.557905 19:-0.498413 20:-0.241071 21:-0.456386 22:-0.729739
0.3316894309565212 1:0.472912 2:-0.71709 3:-1 4:0.957482 5:0.89916 6:-0.676642 7:-1 8:-1 9:-0.724596 10:-0.725191 11:-0.442478 12:-0.63964 13:-0.361072 14:0.863248 15:-0.597271 16:-0.633627 17:-0.582579 18:-0.632128 19:-0.650794 20:-0.441964 21:-0.628258 22:-0.546135
2.000347434232044 1:-0.105208 2:0.43418 3:-1 4:0.971586 5:0.915966 6:-0.269341 7:-0.645796 8:-0.571814 9:-0.467668 10:-0.709924 11:-0.181416 12:-0.464294 13:-0.612643 14:0.811966 15:-0.254796 16:-0.310012 17:-0.135392 18:-0.353473 19:-0.777778 20:-0.209821 21:-0.449494 22:-0.688382
0.01530474389206374 1:-0.727368 2:0.442263 3:0.0909091 4:-0.179716 5:-0.344538 6:-0.641376 7:0.409655 8:0.502471 9:0.271257 10:-0.0941476 11:-0.402655 12:-0.0344979 13:-0.684351 14:-0.333333 15:-0.69821 16:0.548478 17:0.695211 18:0.260279 19:-0.0984127 20:-0.446429 21:-0.106612 22:-0.739641
0.4583882266459824 1:-0.455435 2:0.271363 3:0.434343 4:-0.425905 5:-0.663866 6:0.600894 7:0.0465841 8:0.119679 9:-0.0661153 10:-0.750636 11:-0.349558 12:0.49154 13:-0.925041 14:-0.760684 15:0.383803 16:0.0176072 17:0.310483 18:-0.136695 19:-0.784127 20:-0.491071 21:0.402972 22:-0.950998
0.8302875669034474 1:-0.550507 2:0.218245 3:-0.927273 4:-0.283418 5:-0.478992 6:-0.242022 7:0.315605 8:0.397815 9:0.20475 10:-0.587786 11:-0.132743 12:0.553505 13:-0.792974 14:-0.589744 15:-0.224153 16:0.444436 17:0.559424 18:0.254315 19:-0.28254 20:-0.236607 21:0.571829 22:-0.864449
1.77193794881881 1:-0.660958 2:0.243649 3:-0.999998 4:0.053614 5:-0.0420168 6:-0.61654 7:0.00283798 8:0.120525 9:-0.0322514 10:-0.638677 11:-0.530973 12:-0.414195 13:-0.661818 14:-0.0940171 15:-0.590061 16:0.019363 17:0.290096 18:-0.238997 19:-0.777778 20:-0.517857 21:-0.367219 22:-0.739342
1.027941877282966 1:-0.557497 2:0.700924 3:-0.373737 4:-0.620243 5:-0.915966 6:0.606606 7:0.370467 8:0.358079 9:0.25838 10:-0.155216 11:-0.69469 12:0.21336 13:-0.987548 14:-0.897436 15:0.431441 16:0.533041 17:0.700636 18:0.240322 19:-0.244444 20:-0.696429 21:0.180702 22:-0.979856
0.5088226629248143 1:-0.512758 2:0.512702 3:0.0909091 4:-0.45183 5:-0.697479 6:-0.5773 7:0.13862 8:0.159952 9:-0.0626829 10:-0.791349 11:-0.495575 12:0.264338 13:-0.80533 14:-0.726496 15:-0.569976 16:0.173733 17:0.416124 18:-0.0991678 19:-0.574603 20:-0.522321 21:0.267715 22:-0.868947
1.466204023818776 1:-0.590353 2:0.584296 3:0.434343 4:-0.68516 5:-1 6:0.0731404 7:0.442137 8:0.33904 9:0.442443 10:-0.541985 11:-0.59292 12:0.780268 13:-0.976801 14:-0.982906 15:-0.156946 16:0.360761 17:0.469774 18:-0.0798928 19:-0.631746 20:-0.611607 21:0.682533 22:-0.983134
0.2629708731120864 1:-0.590353 2:0.583141 3:0.878788 4:-0.374054 5:-0.596639 6:-0.82789 7:0.533335 8:0.297158 9:0.815822 10:0.206107 11:-0.588496 12:-0.0678972 13:-0.507196 14:-0.57265 15:-0.892623 16:0.142817 17:0.359879 18:-0.0762007 19:-0.644444 20:-0.602679 21:-0.11135 22:-0.549198
1.369370200334063 1:0.0751485 2:0.406467 3:-0.515152 4:-0.529607 5:-0.798319 6:-0.248479 7:0.346705 8:0.482831 9:0.125849 10:-0.592875 11:-0.389381 12:0.71303 13:-0.885505 14:-0.777778 15:-0.60654 16:0.249373 17:0.499541 18:-0.0384837 19:-0.67619 20:-0.410714 21:0.620504 22:-0.853302
0.3532266105603463 1:-0.631597 2:0.110855 3:-1 4:0.779737 5:0.89916 6:-0.867379 7:0.0114255 8:0.0474588 9:-0.0617383 10:-0.521628 11:-0.566372 12:-0.726214 13:-0.244958 14:0.931624 15:-0.896485 16:-1 17:-1 18:-0.485274 19:-0.434921 20:-0.571429 21:-0.729916 22:-0.308385

0.3384743824264664 1:-0.667948 2:0.495381 3:-0.998323 4:-0.490822 5:-0.747899 6:-0.18043
7:0.306899 8:0.235879 9:0.333664 10:-0.0992366 11:-0.615044 12:0.0863546 13:-0.907256 14:-
0.726496 15:-0.132484 16:0.389641 17:0.568534 18:0.06058 19:-0.530159 20:-0.589286 21:0.120827
22:-0.91499
0.5598021620455723 1:-0.60014 2:0.469977 3:-1 4:0.949808 5:0.529412 6:-1 7:0.460474 8:0.365314
9:0.558179 10:-0.221374 11:-0.995575 12:-1 13:1 14:0.57265 15:-1 16:0.737177 17:0.994021
18:0.302142 19:-0.047619 20:-1 21:-1 22:1
0.5379305525369277 1:-0.444949 2:0.612009 3:-0.777778 4:-0.271181 5:-0.697479 6:-0.918788
7:0.505235 8:0.338113 9:0.68677 10:-0.0178117 11:-1 12:-0.77895 13:-0.419694 14:-0.675214 15:-
0.939488 16:0.87472 17:0.763761 18:0.824177 19:0.263492 20:-1 21:-0.78376 22:-0.466928
0.1479550247338354 1:0.0199231 2:0.136259 3:-0.888889 4:-0.43586 5:-0.428571 6:-0.952068
7:0.205291 8:0.149368 9:0.254049 10:-0.145038 11:-0.995575 12:-0.868161 13:-0.0990532 14:-
0.435897 15:-0.965495 16:0.718257 17:0.769012 18:0.544481 19:0.0412698 20:-1 21:-0.864743 22:-
0.23167
0.18360611330331 1:-0.51835 2:0.180139 3:0.636364 4:0.0527844 5:0.394958 6:-0.532348 7:0.558907
8:0.383524 9:0.750075 10:-0.592875 11:-0.69469 12:-0.725775 13:-0.481434 14:0.299145 15:-0.748938
16:0.388545 17:0.348428 18:0.352261 19:-0.631746 20:-0.700893 21:-0.704932 22:-0.50976
0.1931599389478212 1:-0.612024 2:0.460739 3:0.575758 4:0.26807 5:-0.579832 6:-0.527878 7:0.616122
8:0.590934 9:0.577069 10:0.195929 11:-0.566372 12:-0.045924 13:-0.782643 14:-0.555556 15:-
0.586456 16:0.212515 17:0.281001 18:0.200826 19:-0.130159 20:-0.598214 21:-0.125996 22:-0.809098
0.01298352738544033 1:0.117092 2:0.355658 3:-0.997576 4:0.292336 5:-0.563025 6:-0.618776
7:0.267888 8:0.41487 9:0.0455896 10:-0.267176 11:-0.119469 12:0.727093 13:-0.777911 14:-0.589744
15:-0.6189 16:0.152874 17:0.397697 18:-0.162029 19:-0.612698 20:-0.205357 21:0.629981 22:-
0.846175
0.23577942265077 1:-0.540021 2:0.415704 3:-0.99996 4:0.731619 5:-0.260504 6:0.329691 7:0.574743
8:0.394596 9:0.770324 10:-0.277354 11:-0.606195 12:-0.383872 13:-0.887067 14:-0.316239
15:0.217973 16:-0.14359 17:-0.305058 18:0.119825 19:-0.504762 20:-0.642857 21:-0.393065 22:-
0.915919
1.002779759058828 1:-0.357567 2:0.363741 3:-0.999998 4:0.633931 5:-0.327731 6:-0.324724
7:0.00277917 8:0.282525 9:-0.340137 10:-0.521628 11:-0.561947 12:-0.279719 13:-0.82302 14:-
0.299145 15:-0.203296 16:-0.381634 17:-0.275449 18:-0.421466 19:-0.453968 20:-0.540179 21:-0.2647
22:-0.848437
0.006250757809759752 1:-0.50367 2:0.615473 3:-0.676768 4:0.707145 5:-0.277311 6:-0.366696
7:0.763065 8:0.549621 9:0.96017 10:0.251908 11:-0.553097 12:-0.299495 13:-0.737624 14:-0.350427
15:-0.350586 16:0.0541981 17:0.317601 18:-0.183027 19:-0.326984 20:-0.598214 21:-0.305191 22:-
0.789425
0.1110198471675969 1:-0.641384 2:0.397229 3:-0.994545 4:0.682879 5:-0.294118 6:-0.614802
7:0.939388 8:0.862875 9:0.880555 10:-0.195929 11:-0.588496 12:-0.338167 13:-0.67443 14:-0.316239
15:-0.299858 16:0.124978 17:0.224503 18:0.0836039 19:-0.231746 20:-0.647321 21:-0.399095 22:-
0.783443
1.532949744603857 1:0.380636 2:0.377598 3:-0.996768 4:-0.195686 5:-0.89916 6:0.39625 7:0.274255
8:0.442445 9:-0.0139602 10:-0.796438 11:-0.513274 12:0.679631 13:-0.979278 14:-0.897436
15:0.314278 16:0.138308 17:0.473301 18:-0.2226 19:-0.809524 20:-0.5625 21:0.560629 22:-0.975713
0.09386627656098602 1:-0.641384 2:0.420323 3:-0.998586 4:0.682879 5:-0.294118 6:-0.373401
7:0.911258 8:1 9:0.65415 10:-0.129771 11:-0.513274 12:-0.232257 13:-0.706277 14:-0.316239 15:-
0.311961 16:0.379766 17:0.428223 18:0.246097 19:-0.396825 20:-0.665179 21:-0.426664 22:-0.771746
0.9054342034952781 1:-0.124782 2:0.211316 3:-0.818182 4:0.390024 5:-0.495798 6:-0.27704
7:0.117887 8:0.243797 9:-0.054597 10:-0.720102 11:-0.513274 12:-0.051637 13:0.781593 14:-
0.470085 15:-0.353933 16:-0.0528918 17:0.0643743 18:-0.00302 19:-0.638095 20:-0.53125 21:-
0.103166 22:-0.828002
3.410039009442591 1:-0.667249 2:0.108545 3:-1 4:0.983822 5:0.630252 6:-0.353533 7:-0.121872
8:0.0419146 9:-0.345044 10:-0.720102 11:0.079646 12:-0.165458 13:-0.669901 14:0.538462 15:-
0.374533 16:0.193637 17:0.410888 18:-0.0265173 19:-0.51746 20:0.03125 21:-0.163041 22:-0.726094
1.092712702723601 1:-0.616218 2:0.295612 3:-0.953535 4:-0.582702 5:0.092437 6:-0.598659
7:0.537423 8:0.237684 9:0.908544 10:-0.470738 11:-0.442478 12:-0.383872 13:-0.616666 14:0.0769231
15:-0.704133 16:0.544039 17:0.585363 18:0.403819 19:-0.339683 20:-0.473214 21:-0.404695 22:-
0.672673
2.211720459868876 1:-0.40021 2:-0.184758 3:-1 4:0.0940579 5:0.142857 6:0.0788526 7:0.133106
8:0.181755 9:0.115506 10:-0.613232 11:-0.221239 12:-0.178642 13:-0.777144 14:0.145299 15:0.177546
16:0.0131123 17:0.0949007 18:-0.11174 19:-0.695238 20:-0.138393 21:-0.0954124 22:-0.801365
0.8899710484219431 1:-0.691716 2:0.306005 3:-1 4:0.106087 5:0.159664 6:-0.0254564 7:0.307958
8:0.470133 9:0.0414891 10:-0.603053 11:-0.20354 12:-0.17249 13:-0.770228 14:0.025641 15:-0.468263
16:-0.0758015 17:0.0452202 18:-0.111683 19:-0.657143 20:-0.209821 21:-0.0945509 22:-0.820169
0.09360116547910323 1:-0.510661 2:0.273672 3:0.89899 4:-0.42632 5:-0.092437 6:-0.297901
7:0.0539658 8:0.147807 9:-0.00799373 10:-0.536896 11:-0.0840708 12:0.141727 13:-0.759949 14:-
0.247863 15:-0.404146 16:0.351744 17:0.636214 18:-0.0774314 19:-0.206349 20:-0.09375 21:0.290114
22:-0.803622
0.3819549298880505 1:-0.542817 2:0.338337 3:0.59596 4:-0.318469 5:0.142857 6:-0.56985 7:0.530218
8:0.454118 9:0.58352 10:-0.155216 11:-0.362832 12:-0.325862 13:-0.597505 14:0.0940171 15:-
0.702588 16:1 17:0.821966 18:1 19:-0.28254 20:-0.566964 21:-0.508077 22:-0.639628
0.6410104039785512 1:-0.680531 2:0.375289 3:-0.986465 4:-0.362646 5:0.680672 6:-0.0848131
7:0.0557892 8:0.0952427 9:0.0556106 10:-0.633588 11:0.10177 12:-0.171611 13:-0.653612 14:0.692308
15:-0.231878 16:0.272817 17:0.532662 18:-0.0571339 19:-0.460317 20:0.116071 21:-0.166918 22:-
0.714146
1.889793258756075 1:0.0947221 2:0.215935 3:-1 4:-0.834284 5:-0.243697 6:-0.732522 7:0.233347
8:0.219702 9:0.287199 10:-0.684478 11:-0.420354 12:-0.145682 13:-0.596061 14:-0.333333 15:-
0.753573 16:0.388784 17:0.314738 18:0.412396 19:-0.625397 20:-0.486607 21:-0.157872 22:-0.691433

2.042416189543852 1:-0.0241174 2:-0.192841 3:-1 4:0.122472 5:0.512605 6:-0.775984 7:-0.095124
8:0.149498 9:-0.38211 10:-0.689567 11:0.132743 12:-0.063942 13:-0.471091 14:0.367521 15:-0.721385
16:0.0521193 17:0.288482 18:-0.236195 19:-0.52381 20:0.1875 21:0.0674133 22:-0.644174
0.7698513116920576 1:-0.182803 2:-0.112009 3:-0.998909 4:-0.236545 5:-0.092437 6:-0.892214
7:0.514999 8:0.540744 9:0.435877 10:-0.547074 11:-0.349558 12:-0.168534 13:-0.298438 14:-0.196581
15:-0.938973 16:0.40703 17:0.341658 18:0.422223 19:-0.422222 20:-0.540179 21:-0.330605 22:-
0.361375
0.4239526215419724 1:-0.628102 2:0.241339 3:-0.89697 4:-0.595976 5:-0.462185 6:-0.431019
7:0.345087 8:0.424576 9:0.194384 10:-0.480916 11:-0.287611 12:0.272687 13:-0.779462 14:-0.435897
15:-0.366551 16:0.321993 17:0.35311 18:0.284874 19:-0.644444 20:-0.28125 21:0.252208 22:-0.801086
0.1613095108044116 1:-0.675638 2:-0.189376 3:-0.994747 4:-0.125791 5:0.563025 6:-0.78865 7:-
0.217054 8:-0.00611322 9:-0.367205 10:-0.659033 11:0.181416 12:-0.0507581 13:-0.402116
14:0.401709 15:-0.817175 16:-0.0886259 17:0.250522 18:-0.350765 19:-0.52381 20:0.151786
21:0.0122765 22:-0.537505
1.447523789333504 1:0.0415938 2:0.0461894 3:-0.474747 4:-0.773307 5:-0.848739 6:-0.68906
7:0.713981 8:0.586024 9:0.809348 10:-0.628499 11:-0.765487 12:-0.106131 13:-0.733468 14:-0.897436
15:-0.722158 16:0.864368 17:0.777284 18:0.790645 19:-0.739683 20:-0.825893 21:-0.174241 22:-
0.764443
0.726818603805476 1:-0.63719 2:0.730947 3:1 4:-0.688479 5:-0.663866 6:-0.825903 7:0.720657
8:0.578074 9:0.830749 10:-0.470738 11:-0.663717 12:-0.126346 13:-0.523373 14:-0.641026 15:-
0.848075 16:0.871939 17:0.964175 18:0.577862 19:1 20:-0.830357 21:-0.470601 22:-0.574144
1.690866432070884 1:0.380636 2:0.367206 3:0.252525 4:-0.796536 5:-0.89916 6:-0.57283 7:0.414243
8:0.480652 9:0.302702 10:-0.450382 11:-0.743363 12:0.0503186 13:-0.845293 14:-0.880342 15:-
0.511266 16:0.569168 17:0.800471 18:0.170815 19:-0.244444 20:-0.741071 21:0.0299375 22:-0.862917
0.9768778064499603 1:-0.202377 2:-0.287529 3:-1 4:-0.35435 5:0.915966 6:-0.653794 7:-0.168309 8:-
0.322001 9:0.157248 10:-0.597964 11:0.146018 12:0.238409 13:-0.544238 14:0.931624 15:-0.693575
16:-0.329368 17:-0.636467 18:0.126338 19:-0.352381 20:0.116071 21:-0.264269 22:-0.586299
2.647768485392427 1:-0.691716 2:0.555427 3:-0.80404 4:-0.776003 5:-0.731092 6:-0.557184
7:0.422463 8:0.222482 9:0.683315 10:-0.506361 11:-0.548673 12:0.211162 13:-0.824193 14:-0.709402
15:-0.365263 16:0.415514 17:0.402743 18:0.416997 19:-0.320635 20:-0.535714 21:0.209994 22:-
0.869956
3.190340074690057 1:-0.0590703 2:0.146651 3:-0.977778 4:-0.433994 5:0.310924 6:-0.00707811 7:-
0.0843455 8:-0.134296 9:0.0946348 10:-0.806616 11:-0.141593 12:-0.198857 13:-0.756816 14:0.230769
15:-0.0245912 16:-0.32028 17:-0.107127 18:-0.362429 19:-0.638095 20:-0.142857 21:-0.153134 22:-
0.812974
0.8590973527764005 1:-0.590353 2:0.372979 3:-0.69697 4:-0.682879 5:-0.394958 6:-0.115112
7:0.142987 8:0.0750658 9:0.308415 10:-0.826972 11:-0.424779 12:-0.0177983 13:-0.860404 14:-
0.367521 15:-0.219003 16:0.188454 17:0.355893 18:0.0223139 19:-0.746032 20:-0.428571 21:-
0.0445832 22:-0.868078
1.062505421238179 1:-0.540021 2:0.255196 3:-0.971717 4:-0.202945 5:0.966387 6:0.218925 7:-
0.0510837 8:-0.02364 9:-0.037665 10:-0.745547 11:0.172566 12:-0.23797 13:-0.722236 14:0.82906
15:0.197116 16:-0.530161 17:-0.307162 18:-0.602533 19:-0.612698 20:0.138393 21:-0.208701 22:-
0.785708
0.07588237504889241 1:-0.65187 2:0.331409 3:0.515152 4:-0.398527 5:0.411765 6:-0.746182
7:0.165162 8:-0.0391019 9:0.451957 10:-0.694656 11:-0.318584 12:-0.400132 13:-0.49461 14:0.452991
15:-0.758208 16:0.00212803 17:-0.0641529 18:0.110869 19:-0.396825 20:-0.308036 21:-0.404695 22:-
0.600932
3.6 1:0.143656 2:-0.431871 3:-0.996566 4:-0.260396 5:0.210084 6:-0.399478 7:0.171264 8:0.208841
9:0.153977 10:-0.450382 11:-0.247788 12:-0.243683 13:-0.629715 14:0.042735 15:-0.429638
16:0.200449 17:0.178223 18:0.238221 19:-0.466667 20:-0.232143 21:-0.12815 22:-0.6919
0.4111348127090822 1:0.17791 2:0.230947 3:0.111111 4:-0.963912 5:-0.596639 6:-0.629703 7:0.327795
8:0.414886 9:0.220761 10:-0.776081 11:-0.358407 12:0.354428 13:-0.764398 14:-0.709402 15:-
0.698725 16:0.554756 17:0.555185 18:0.427998 19:-0.784127 20:-0.714286 21:-0.167779 22:-0.760798
1.876893338755907 1:-0.272981 2:-0.215935 3:-1 4:-0.583117 5:0.747899 6:0.160065 7:-0.0357321
8:0.15657 9:-0.226681 10:-0.577608 11:-0.0176991 12:-0.290705 13:-0.689562 14:0.709402
15:0.0668212 16:-0.270137 17:0.247169 18:-0.535961 19:-0.663492 20:-0.0491071 21:-0.29916 22:-
0.743067
0.1493771364550289 1:-0.632296 2:0.19746 3:-0.575758 4:0.414912 5:-0.563025 6:-0.211226
7:0.693953 8:0.649351 9:0.667281 10:-0.338422 11:-0.712389 12:-0.325423 13:-0.786628 14:-0.57265
15:-0.466976 16:0.605787 17:0.762764 18:0.309962 19:-0.574603 20:-0.65625 21:-0.205686 22:-
0.817034
0.0612882885863129 1:-0.632296 2:0.19746 3:-0.575758 4:0.414912 5:-0.563025 6:-0.287719
7:0.618695 8:0.632036 9:0.531411 10:-0.51145 11:-0.566372 12:-0.0648209 13:-0.817197 14:-0.57265
15:-0.532638 16:0.625888 17:0.78331 18:0.336186 19:-0.619048 20:-0.700893 21:-0.286237 22:-
0.811818
1.534528187915932 1:-0.387627 2:0.369515 3:-0.980404 4:-0.651976 5:-0.731092 6:-0.494102
7:0.281225 8:0.145254 9:0.460803 10:0.0636132 11:-0.60177 12:0.0916282 13:-0.842964 14:-0.709402
15:-0.574353 16:0.110398 17:0.0667784 18:0.165021 19:-0.574603 20:-0.566964 21:0.138919 22:-
0.837709
1.358258482273513 1:-0.540021 2:0.258661 3:-0.999986 4:-0.440838 5:-0.344538 6:-0.561654 7:-
0.0145428 8:0.239115 9:-0.287429 10:-0.582697 11:-0.579646 12:-0.291145 13:-0.72807 14:-0.333333
15:-0.783443 16:-0.0636795 17:0.0579369 18:-0.0656354 19:-0.536508 20:-0.589286 21:-0.306483 22:-
0.738453
0.9037156266016728 1:-0.602936 2:0.312933 3:-1 4:-0.183864 5:0.12605 6:-0.394511 7:-0.0632885
8:0.105291 9:-0.166025 10:-0.577608 11:-0.464602 12:-0.419468 13:-0.662511 14:0.00854701 15:-
0.460281 16:-0.599185 17:-0.620698 18:-0.198819 19:-0.542857 20:-0.504464 21:-0.402111 22:-
0.756256

1.530365206792466 1:-0.423978 2:0.474596 3:-0.616162 4:-0.688686 5:-0.798319 6:-0.431765
7:0.468473 8:0.527997 9:0.344076 10:-0.328244 11:-0.323009 12:0.874753 13:-0.856765 14:-0.777778
15:-0.6292 16:0.33139 17:0.464128 18:0.0985809 19:-0.701587 20:-0.285714 21:0.913418 22:-0.827559
0.4075113295209083 1:-0.497379 2:0.407621 3:-0.999986 4:-0.615265 5:-0.663866 6:-0.350801
7:0.450181 8:0.553751 9:0.27158 10:-0.338422 11:-0.584071 12:0.0309822 13:-0.842671 14:-0.675214
15:-0.552981 16:0.2721 17:0.421359 18:0.0461332 19:-0.593651 20:-0.674107 21:-0.132027 22:-
0.783008
3.6 1:-0.663754 2:0.264434 3:-1 4:0.795499 5:0.647059 6:-0.137961 7:-0.198791 8:-0.205638 9:-
0.0963164 10:-0.78117 11:-0.00884956 12:-0.242364 13:-0.700858 14:0.65812 15:-0.257113 16:-
0.311824 17:-0.158358 18:-0.313882 19:-0.606349 20:0.0133929 21:-0.231101 22:-0.74152
0.9987061070328201 1:0.150647 2:0.124711 3:0.171717 4:-0.712952 5:-0.579832 6:-0.434496
7:0.386642 8:0.435714 9:0.248888 10:-0.684478 11:-0.115044 12:0.768403 13:-0.800045 14:-0.555556
15:-0.534956 16:0.357011 17:0.458465 18:0.226443 19:-0.644444 20:-0.0848214 21:0.779453 22:-
0.830602
0.1763577735703556 1:-0.605033 2:0.367206 3:0.151515 4:0.0256144 5:-0.176471 6:-0.912828
7:0.732479 8:0.730856 9:0.658458 10:-0.00254453 11:-0.659292 12:-0.496374 13:-0.284942 14:-
0.145299 15:-0.88979 16:0.513249 17:0.723649 18:0.191207 19:-0.206349 20:-0.651786 21:-0.499031
22:-0.427346
1.634970408785411 1:-0.465222 2:0.311778 3:-0.868687 4:-0.155242 5:-0.394958 6:-0.53657
7:0.273858 8:0.230368 9:0.344721 10:-0.430025 11:0.115044 12:0.793891 13:-0.723546 14:-0.384615
15:-0.477276 16:0.149728 17:0.406602 18:-0.118973 19:-0.619048 20:0.129464 21:0.80056 22:-
0.789812
0.1629851814501007 1:-0.689619 2:0.229792 3:0.89899 4:-0.46303 5:0.495798 6:-0.0793493
7:0.0887569 8:0.0982343 9:0.175885 10:-0.669211 11:0.10177 12:-0.0823995 13:-0.744497 14:0.452991
15:-0.0727437 16:-0.103642 17:0.0496172 18:-0.258972 19:-0.68254 20:0.0982143 21:-0.0618135 22:-
0.792759
2.040396661379445 1:-0.403006 2:0.221709 3:-0.993333 4:-0.596599 5:0.0588235 6:0.0676766
7:0.153913 8:0.158505 9:0.194591 10:-0.689567 11:0.0707965 12:0.184355 13:-0.804286 14:0.0769231
15:0.00708124 16:-0.230242 17:-0.141908 18:-0.191472 19:-0.75873 20:0.0982143 21:0.197502 22:-
0.83063
0.1453422070784811 1:-0.623209 2:0.357968 3:-0.010101 4:-0.592036 5:1 6:-0.861418 7:0.420242
8:0.279469 9:0.60112 10:-0.430025 11:0.314159 12:-0.15579 13:-0.286732 14:0.880342 15:-0.8661
16:-0.083513 17:0.206409 18:-0.256019 19:-0.549206 20:0.290179 21:-0.123842 22:-0.447223
0.9219972332135861 1:-0.698008 2:0.218245 3:-1 4:0.251685 5:0.579832 6:-0.332174 7:0.128077
8:0.111257 9:0.128729 10:-0.440204 11:0.261062 12:0.00329598 13:-0.686616 14:0.487179 15:-
0.0552337 16:-0.101773 17:0.0561812 18:-0.258688 19:-0.68254 20:0.28125 21:0.0704286 22:-0.757165
0.4905797693674867 1:-0.65187 2:0.43418 3:-0.616162 4:-0.587888 5:-0.495798 6:-0.559916
7:0.120195 8:0.216727 9:0.0440922 10:-0.577608 11:-0.517699 12:-0.05735 13:-0.774101 14:-0.538462
15:-0.429123 16:0.23985 17:0.143648 18:0.368469 19:-0.580952 20:-0.455357 21:0.107474 22:-
0.829501
1.633299026550503 1:-0.667948 2:0.596998 3:0.131313 4:-0.856061 5:-0.747899 6:-0.287471
7:0.299547 8:0.271372 9:0.320671 10:-0.694656 11:-0.59292 12:0.135575 13:-0.896791 14:-0.726496
15:-0.292391 16:0.0428065 17:0.326474 18:-0.0839826 19:-0.67619 20:-0.616071 21:0.0631058 22:-
0.91174
0.9697028998462834 1:0.441454 2:-0.163972 3:-0.945455 4:-0.181375 5:0.781513 6:-0.146157
7:0.186219 8:0.357038 9:-0.0493446 10:-0.755725 11:0.530973 12:0.0986596 13:-0.733372 14:0.299145
15:0.489378 16:-0.306458 17:0.048684 18:-0.529353 19:-0.752381 20:0.169643 21:0.0945509 22:-
0.822239
0.692350945688763 1:0.142957 2:-0.244804 3:0.494949 4:-0.240485 5:-0.092437 6:-0.196076
7:0.0988883 8:0.249276 9:-0.0474556 10:-0.56743 11:0.123894 12:0.388706 13:-0.763482 14:-0.316239
15:-0.389468 16:-0.0343084 17:0.360638 18:-0.367617 19:-0.695238 20:0.0178571 21:0.533061 22:-
0.82919
3.6 1:-0.182803 2:0.0508083 3:-0.999996 4:-0.260396 5:-0.12605 6:-0.517695 7:-0.139061 8:-
0.0733099 9:-0.0962473 10:-0.720102 11:0.0884956 12:0.382993 13:-0.726647 14:-0.179487 15:-
0.437363 16:-0.0647329 17:0.300139 18:-0.36739 19:-0.6 20:0.0892857 21:0.448632 22:-0.799116
3.6 1:0.230339 2:-0.294457 3:-0.999939 4:-0.131391 5:0.092437 6:-0.630945 7:-0.0685675 8:-
0.0314929 9:-0.111405 10:-0.557252 11:0.110619 12:0.198418 13:-0.670711 14:-0.128205 15:-0.277198
16:-0.0487481 17:0.123529 18:-0.166763 19:-0.574603 20:-0.0133929 21:0.260392 22:-0.810139
1.815716717985824 1:0.428871 2:0.0323326 3:-0.373737 4:-0.704034 5:-0.697479 6:-0.675152
7:0.079096 8:0.0452151 9:0.128038 10:-0.709924 11:-0.823009 12:-0.412876 13:-0.692946 14:-
0.675214 15:-0.671173 16:0.516438 17:0.491459 18:0.492488 19:-0.669841 20:-0.799107 21:-0.384019
22:-0.755737
0.07830506361364406 1:-0.223348 2:0.100462 3:-1 4:-0.109821 5:-0.0588235 6:0.0356389 7:0.22648
8:0.201346 9:0.262204 10:-0.399491 11:-0.190265 12:-0.0107669 13:-0.84444 14:-0.264957
15:0.0137762 16:-0.142396 17:-0.0479248 18:-0.217241 19:-0.650794 20:-0.25 21:0.100151 22:-
0.880593
1.254710240233339 1:-0.475708 2:0.209007 3:-1 4:-0.12102 5:-0.0756303 6:0.412641 7:0.161089
8:0.226693 9:0.120044 10:-0.414758 11:-0.0663717 12:0.148319 13:-0.840812 14:-0.282051 15:-
0.151539 16:-0.16015 17:-0.0938409 18:-0.104413 19:-0.561905 20:-0.522321 21:-0.2479 22:-0.848628
1.329333526752587 1:-0.717581 2:0.357968 3:-1 4:0.181582 5:0.378151 6:0.174966 7:0.0342323
8:0.209069 9:-0.140155 10:-0.603053 11:0.300885 12:0.162821 13:-0.752634 14:0.230769 15:0.254538
16:-0.298929 17:-0.0228711 18:-0.560291 19:-0.67619 20:-0.133929 21:-0.144088 22:-0.829972
0.1627024199353362 1:-0.472912 2:0.183603 3:-0.212121 4:-0.763144 5:0.327731 6:-0.605861
7:0.042908 8:-0.051117 9:0.182151 10:-0.664122 11:-0.079646 12:-0.149637 13:-0.654214 14:0.350427
15:-0.614523 16:-0.324985 17:-0.249241 18:-0.216541 19:-0.720635 20:-0.0446429 21:-0.129011 22:-
0.721316

2.31835817113451 1:0.43726 2:-0.590069 3:-1 4:0.591621 5:0.210084 6:-0.153856 7:-0.0322766 8:-0.139547 9:0.169251 10:-0.613232 11:-0.0884956 12:-0.0916282 13:-0.763946 14:0.230769 15:-0.0838161 16:-0.0142501 17:-0.0489371 18:0.0253813 19:-0.288889 20:-0.0446429 21:-0.0600905 22:-0.802055
0.866405904362975 1:0.43726 2:-0.590069 3:-1 4:0.591621 5:0.210084 6:-0.212716 7:0.00817575 8:-0.171869 9:0.308507 10:-0.6743 11:-0.0442478 12:-0.0476818 13:-0.782136 14:0.230769 15:-0.123729 16:0.146792 17:0.218651 18:0.164074 19:-0.428571 20:-0.236607 21:-0.242731 22:-0.839523
0.0566635537335906 1:-0.566585 2:0.174365 3:-0.515152 4:-0.0776729 5:0.966387 6:-0.767292 7:0.471561 8:0.628524 9:0.189408 10:-0.592875 11:-0.40708 12:-0.628214 13:-0.412112 14:1 15:-0.748165 16:0.0226218 17:0.363691 18:-0.335523 19:-0.320635 20:-0.397321 21:-0.625673 22:-0.529018
0.5877958601653774 1:-0.525341 2:0.0300231 3:-1 4:0.983615 5:0.596639 6:-0.367689 7:-0.00435256 8:0.0409391 9:0.00686494 10:-0.475827 11:0.234513 12:-0.0287849 13:-0.611295 14:0.57265 15:-0.764903 16:0.0242933 17:0.377657 18:-0.295932 19:-0.542857 20:0.174107 21:-0.0648288 22:-0.650204
0.9017889401754576 1:0.407899 2:0.0473441 3:-0.272727 4:-0.323032 5:-0.798319 6:-0.0540171 7:0.538923 8:0.461679 9:0.570942 10:-0.821883 11:-0.615044 12:0.178203 13:-0.88566 14:-0.846154 15:0.0650187 16:0.51832 17:0.810436 18:0.128876 19:-0.733333 20:-0.709821 21:0.0488908 22:-0.907871
0.2666050545666639 1:0.0415938 2:0.297921 3:0.979798 4:-0.731411 5:-0.865546 6:-0.700236 7:0.517131 8:0.512926 9:0.485268 10:-0.455471 11:-0.792035 12:-0.13953 13:-0.82399 14:-0.897436 15:-0.235483 16:0.697946 17:0.782424 18:0.466321 19:-0.498413 20:-0.839286 21:-0.204394 22:-0.893721
0.1984471024158463 1:-0.710591 2:0.312933 3:-1 4:-0.200249 5:0.294118 6:-0.0930088 7:0.322177 8:0.169512 9:0.537262 10:-0.608142 11:-0.29646 12:-0.333333 13:-0.752239 14:0.316239 15:-0.00424874 16:0.345591 17:0.445796 18:0.167823 19:-0.409524 20:-0.308036 21:-0.351281 22:-0.809628
2.067373186536643 1:-0.532331 2:0.553118 3:-0.737374 4:-0.948771 5:-0.831933 6:-0.576804 7:0.543026 8:0.427487 9:0.634385 10:-0.430025 11:-0.853982 12:-0.349154 13:-0.798116 14:-0.811966 15:-0.575383 16:0.591923 17:0.850452 18:0.179657 19:-0.28254 20:-0.852679 21:-0.35602 22:-0.823702
0.1043294465824329 1:-0.655365 2:0.0635104 3:-1 4:0.0349476 5:0.310924 6:-0.756116 7:0.533659 8:0.459581 9:0.591675 10:-0.358779 11:-0.0221239 12:-0.0894309 13:-0.538553 14:0.196581 15:-0.779065 16:0.184437 17:0.477825 18:-0.12329 19:-0.473016 20:0.0401786 21:0.043291 22:-0.635321
1.207861025010369 1:-0.667948 2:0.277136 3:-0.999919 4:-0.129524 5:0.0420168 6:-0.711412 7:0.201747 8:0.347966 9:0.0209173 10:-0.389313 11:0.207965 12:0.346078 13:-0.698295 14:0.025641 15:-0.71675 16:-0.318144 17:-0.118357 18:-0.277017 19:-0.498413 20:0.15625 21:0.30476 22:-0.743394
3.181694578671059 1:-0.507165 2:0.325635 3:-0.737374 4:-0.632272 5:-0.512605 6:-0.27853 7:0.010205 8:-0.0867558 9:0.116658 10:-0.720102 11:-0.769912 12:-0.466491 13:-0.774581 14:-0.538462 15:-0.414446 16:0.291217 17:0.373529 18:0.179013 19:-0.422222 20:-0.8125 21:-0.508938 22:-0.781556
0.2381919424163903 1:-0.38483 2:0.329099 3:-0.535354 4:-0.489785 5:-0.495798 6:-0.564138 7:0.6314 8:0.533021 9:0.694925 10:-0.750636 11:-0.561947 12:-0.132498 13:-0.780415 14:-0.641026 15:-0.789365 16:0.483373 17:0.565244 18:0.335864 19:-0.574603 20:-0.486607 21:0.188886 22:-0.747614
0.6581595378934142 1:-0.628102 2:0.354503 3:-0.79798 4:-0.799855 5:-0.462185 6:-0.759096 7:0.647472 8:0.640458 9:0.578405 10:-0.358779 11:-0.433628 12:0.0402109 13:-0.546652 14:-0.470085 15:-0.69924 16:0.666861 17:0.654229 18:0.587556 19:-0.638095 20:-0.544643 21:-0.120396 22:-0.648825
0.2745311355071652 1:0.322614 2:-0.490762 3:-0.59596 4:-0.618791 5:0.613445 6:-0.571092 7:0.505735 8:0.342145 9:0.694948 10:-0.379135 11:0.137168 12:-0.113601 13:-0.525701 14:0.0940171 15:-0.801983 16:0.260161 17:0.237046 18:0.320641 19:-0.0793651 20:-0.0535714 21:0.0230454 22:-0.558435
0.5330673491038334 1:-0.529535 2:0.39261 3:-0.99897 4:0.26247 5:-0.529412 6:-0.551968 7:0.188498 8:0.282753 9:0.0397844 10:-0.363868 11:-0.19469 12:0.536366 13:-0.828892 14:-0.521368 15:-0.592636 16:0.0545633 17:0.315228 18:-0.203988 19:-0.68254 20:-0.196429 21:0.520569 22:-0.859615
1.548778091420242 1:-0.00803915 2:-0.0242494 3:-0.999737 4:0.746552 5:-0.176471 6:-0.423321 7:-0.00117637 8:0.117224 9:-0.0645719 10:-0.735369 11:-0.50885 12:-0.306966 13:-0.717824 14:-0.196581 15:-0.377881 16:-0.00316747 17:0.184977 18:-0.149362 19:-0.822222 20:-0.486607 21:-0.265561 22:-0.78365
0.3733020936612817 1:0.0367005 2:0.319861 3:0.494949 4:-0.0830654 5:-0.781513 6:0.20303 7:0.0465841 8:0.259064 9:-0.201778 10:-0.424936 11:-0.539823 12:0.321907 13:-0.952551 14:-0.82906 15:0.249388 16:0.234681 17:0.528802 18:-0.106041 19:-0.701587 20:-0.629464 21:0.219901 22:-0.976099
0.1430274140805344 1:-0.142258 2:0.277136 3:-0.999996 4:0.124339 5:-0.630252 6:-0.669688 7:0.365821 8:0.353429 9:0.338179 10:-0.562341 11:-0.646018 12:-0.136893 13:-0.684005 14:-0.641026 15:-0.527746 16:0.239232 17:0.381991 18:0.0901173 19:-0.530159 20:-0.625 21:-0.0747362 22:-0.78493
1.628371127446055 1:-0.66655 2:0.468822 3:-0.996162 4:0.792596 5:-0.142857 6:-0.109897 7:0.313428 8:0.350161 9:0.273261 10:-0.455471 11:-0.261062 12:-0.0243902 13:-0.847008 14:-0.179487 15:-0.0436462 16:-0.174211 17:-0.0959129 18:-0.196073 19:-0.580952 20:-0.276786 21:-0.0105535 22:-0.886571
1.052523347932883 1:-0.0367005 2:0.292148 3:-0.923232 4:0.446853 5:-0.394958 6:-0.172234 7:0.211393 8:0.299207 9:0.0842913 10:-0.653944 11:-0.225664 12:0.282356 13:-0.817245 14:-0.452991 15:-0.242693 16:0.294898 17:0.590472 18:-0.0195117 19:-0.155556 20:-0.245536 21:0.335774 22:-0.85922

1.278237524167965 1:-0.47431 2:0.382217 3:-0.980202 4:0.308721 5:-0.495798 6:-0.125047 7:0.196453
8:0.337235 9:0.00801677 10:-0.613232 11:-0.455752 12:0.0410899 13:-0.881546 14:-0.504274 15:-
0.215398 16:0.160319 17:0.652521 18:-0.291653 19:-0.663492 20:-0.473214 21:0.0338143 22:-0.891101
0.1923871405567576 1:-0.396714 2:0.21709 3:-0.59596 4:-0.527118 5:-0.210084 6:-0.715386
7:0.520189 8:0.345576 9:0.720887 10:-0.56743 11:-0.336283 12:-0.0630631 13:-0.515369 14:-0.264957
15:-0.738123 16:0.382983 17:0.326775 18:0.386021 19:-0.453968 20:-0.308036 21:0.0251992 22:-
0.649132
0.5718005890330268 1:0.428871 2:0.0889145 3:-0.414141 4:-0.558021 5:-0.596639 6:-0.628958
7:0.344823 8:0.522209 9:0.0942662 10:-0.664122 11:-0.716814 12:-0.30301 13:-0.781193 14:-0.675214
15:-0.529548 16:0.440194 17:0.575051 18:0.218359 19:0.0031746 20:-0.75 21:-0.281499 22:-0.836815
2.539199525370392 1:-0.31143 2:0.103926 3:-1 4:0.0824432 5:0.529412 6:-0.071402 7:-0.0540982
8:0.0105356 9:-0.0586284 10:-0.760814 11:0.792035 12:0.465172 13:-0.676236 14:0.57265 15:-
0.207416 16:-0.217937 17:0.0121315 18:-0.257628 19:-0.720635 20:0.825893 21:0.456817 22:-0.738301
2.737359314557267 1:-0.215659 2:0.258661 3:-0.737374 4:-0.622524 5:-0.512605 6:-0.0110518
7:0.178234 8:0.0419959 9:0.390311 10:-0.770992 11:-0.575221 12:-0.135575 13:-0.852038 14:-
0.487179 15:0.00321875 16:-0.110271 17:-0.00620018 18:-0.126282 19:-0.822222 20:-0.571429 21:-
0.15098 22:-0.880413
2.507272186582642 1:-0.146452 2:-0.168591 3:-1 4:-0.524837 5:0.680672 6:-0.535825 7:-0.371688 8:-
0.353315 9:-0.282729 10:-0.430025 11:0.0353982 12:-0.223467 13:-0.611657 14:0.641026 15:-0.655208
16:-0.10374 17:0.309265 18:-0.446648 19:-0.447619 20:0.00892857 21:-0.225932 22:-0.674683
1.160436234550747 1:0.165327 2:-0.37067 3:-1 4:-0.444571 5:1 6:-0.681113 7:-0.513572 8:-0.438721
9:-0.491004 10:-0.664122 11:-0.39823 12:-0.63085 13:-0.411649 14:0.74359 15:-0.646453 16:-
0.435137 17:0.0351923 18:-0.754916 19:-0.663492 20:-0.482143 21:-0.631273 22:-0.577003
3.048296730599391 1:-0.247815 2:-0.0819861 3:-1 4:0.903142 5:0.210084 6:-0.72532 7:-0.144428
8:0.0380776 9:-0.308323 10:-0.694656 11:-0.0132743 12:-0.0164799 13:-0.562786 14:0.247863 15:-
0.70851 16:-0.346729 17:-0.0248323 18:-0.587159 19:-0.371429 20:-0.0535714 21:-0.0799052 22:-
0.655808
0.955756145785474 1:0.341489 2:-0.0207852 3:-0.949495 4:-0.605932 5:-0.647059 6:-0.130014
7:0.661545 8:0.72523 9:0.474118 10:-0.78117 11:-0.641593 12:-0.104812 13:-0.86814 14:-0.709402
15:-0.182181 16:0.306191 17:0.491601 18:0.129444 19:-0.726984 20:-0.763393 21:-0.268576 22:-
0.864429
0.01541098899024072 1:-0.540021 2:0.187067 3:-0.999999 4:-0.0187701 5:0.428571 6:-0.424562 7:-
0.123768 8:-0.175934 9:0.11857 10:-0.715013 11:-0.345133 12:-0.431334 13:-0.670263 14:0.435897
15:-0.67864 16:-0.115328 17:-0.13884 18:-0.0577777 19:-0.752381 20:-0.3125 21:-0.402541 22:-
0.658316
0.2552309627201203 1:-0.498078 2:0.0808314 3:-1 4:0.201493 5:0.831933 6:-0.586986 7:-0.297165 8:-
0.314977 9:-0.174411 10:-0.704835 11:-0.247788 12:-0.485827 13:-0.634201 14:0.82906 15:-0.665508
16:-0.238642 17:-0.333813 18:-0.0806123 19:-0.688889 20:-0.361607 21:-0.563213 22:-0.641104
-2.379478785087458 1:0.324013 2:-0.372979 3:-1 4:0.504304 5:0.529412 6:-0.4792 7:-0.508911 8:-
0.383686 9:-0.585616 10:-0.78626 11:-0.495575 12:-0.588662 13:-0.656927 14:0.555556 15:-0.488863
16:-0.763065 17:-0.456709 18:-0.731134 19:-0.587302 20:-0.504464 21:-0.60112 22:-0.728228
-1.085517444937282 1:0.51136 2:-0.398383 3:-1 4:0.373846 5:0.361345 6:-0.0845648 7:-0.594153 8:-
0.508926 9:-0.398973 10:-0.750636 11:-0.486726 12:-0.530213 13:-0.712581 14:0.401709 15:-
0.0925711 16:-0.407634 17:-0.224124 18:-0.30059 19:-0.485714 20:-0.473214 21:-0.527461 22:-
0.777165
-0.2340858449864472 1:-0.191192 2:-0.293303 3:-1 4:0.903764 5:0.915966 6:-0.212219 7:-0.598138
8:-0.233424 9:-1 10:-0.745547 11:0.309735 12:-0.12283 13:-0.653207 14:0.931624 15:-0.187331 16:-
0.534909 17:-0.20062 18:-0.747285 19:-0.714286 20:0.321429 21:-0.123842 22:-0.73086
-3.6 1:-0.34079 2:0.0635104 3:-0.999982 4:0.0934357 5:-0.0756303 6:-0.485409 7:0.0394524
8:0.36068 9:-0.314036 10:-0.664122 11:-0.433628 12:-0.279719 13:-0.688092 14:-0.111111 15:-
0.533411 16:-0.156695 17:0.177243 18:-0.440892 19:-0.733333 20:-0.383929 21:-0.197932 22:-
0.749492
-2.707268111740729 1:-0.193289 2:-0.220554 3:-1 4:-0.276366 5:-0.0756303 6:-0.584503 7:0.0477311
8:-0.0241928 9:0.184432 10:-0.531807 11:-0.283186 12:-0.104812 13:-0.716673 14:-0.0769231 15:-
0.465946 16:0.22718 17:0.243673 18:0.1924 19:-0.638095 20:-0.34375 21:-0.174241 22:-0.784248
-2.908708996206382 1:0.254107 2:-0.180139 3:-1 4:-0.32324 5:-0.159664 6:-0.498572 7:0.224054
8:0.115111 9:0.37642 10:-0.51145 11:-0.309735 12:-0.0709734 13:-0.752404 14:-0.128205 15:-
0.464143 16:0.0176493 17:0.0948532 18:-0.0104801 19:-0.574603 20:-0.28125 21:-0.0622442 22:-
0.797632
-1.72136090167138 1:0.254107 2:-0.165127 3:-1 4:-0.32324 5:-0.159664 6:-0.544766 7:0.138811
8:0.103242 9:0.193531 10:-0.628499 11:-0.29646 12:-0.0577895 13:-0.750939 14:-0.128205 15:-
0.490666 16:0.15876 17:0.192933 18:0.135351 19:-0.574603 20:-0.285714 21:-0.064398 22:-0.803547
-1.337371505638952 1:0.557497 2:-0.991917 3:-1 4:0.23053 5:0.831933 6:-0.251211 7:-0.0906391 8:-
0.159724 9:0.0636043 10:-0.603053 11:0.384956 12:-0.0323006 13:-0.63418 14:0.846154 15:-0.254538
16:-0.213583 17:-0.308443 18:0.0139261 19:-0.485714 20:0.410714 21:-0.0243377 22:-0.710859
-3.6 1:0.150647 2:-0.32448 3:-1 4:-0.248159 5:-0.0252101 6:-0.678381 7:0.147325 8:0.0756999
9:0.297381 10:-0.379135 11:-0.29646 12:-0.154032 13:-0.683483 14:0.00854701 15:-0.536243
16:0.123953 17:0.149136 18:0.127096 19:-0.492063 20:-0.276786 21:-0.154426 22:-0.766665
-1.77692082238554 1:0.140161 2:-0.506928 3:-1 4:-0.149642 5:0.495798 6:0.758599 7:-0.0356586
8:0.126362 9:-0.254809 10:-0.582697 11:0.146018 12:-0.0454845 13:-0.763365 14:0.470085 15:0.6086
16:0.150852 17:0.467481 18:-0.216749 19:-0.409524 20:0.129464 21:-0.0441525 22:-0.816344
-1.162343662732361 1:-0.49598 2:-0.224018 3:-1 4:0.420927 5:0.361345 6:0.350056 7:-0.154501 8:-
0.166407 9:-0.139188 10:-0.613232 11:0.323009 12:0.192265 13:-0.790763 14:0.316239 15:0.271533
16:-0.0991607 17:0.268601 18:-0.345634 19:-0.587302 20:0.428571 21:0.318975 22:-0.816903
-0.7560597042167041 1:-0.170919 2:0.114319 3:-1 4:-0.131391 5:-0.243697 6:-0.21967 7:-0.223451
8:-0.112331 9:-0.262703 10:-0.740458 11:-0.548673 12:-0.31971 13:-0.797834 14:-0.213675 15:-

0.180379 16:0.0683007 17:0.220075 18:-0.0750646 19:-0.0619048 20:-0.0540179 21:-0.323713 22:-0.843759
-0.7574344753330655 1:0.0290108 2:0.211316 3:-0.993939 4:-0.395209 5:-0.596639 6:-0.131255
7:0.034791 8:0.104721 9:-0.0200189 10:-0.745547 11:-0.50885 12:0.0815205 13:-0.88939 14:-0.692308
15:0.00012875 16:0.345184 17:0.570859 18:0.0503744 19:-0.68254 20:-0.566964 21:0.116519 22:-0.938248
-0.8624764597566472 1:-0.182803 2:0.528868 3:-0.414141 4:-0.571295 5:-0.831933 6:-0.0065814
7:0.160089 8:0.259487 9:0.00412357 10:-0.419847 11:-0.876106 12:-0.397495 13:-0.871268 14:-0.811966
15:0.0274237 16:0.468722 17:0.739608 18:0.0862737 19:-0.346032 20:-0.879464 21:-0.420633
22:-0.89201
-1.570025009725839 1:-0.289759 2:0.290993 3:-0.999998 4:-0.294618 5:-0.462185 6:0.0614678 7:-0.0809929
8:-0.0259975 9:0.00398535 10:-0.562341 11:-0.597345 12:-0.224346 13:-0.828802 14:-0.435897
15:0.107506 16:0.0371177 17:0.195796 18:-0.0850429 19:-0.339683 20:-0.602679 21:-0.248762
22:-0.864816
-1.91031414914683 1:-0.521846 2:0.625866 3:-0.171717 4:-0.750493 5:-0.932773 6:-0.353284
7:0.520425 8:0.369948 9:0.693382 10:-0.699746 11:-0.473451 12:0.902878 13:-0.957352 14:-0.91453
15:-0.346208 16:0.496365 17:0.476386 18:0.363092 19:-0.536508 20:-0.459821 21:0.894465 22:-0.953746
-0.6316710677419466 1:0.528137 2:-0.646651 3:-1 4:0.959764 5:0.193277 6:0.460822 7:-0.440652 8:-0.205736
9:-0.511599 10:-0.745547 11:0.247788 12:0.250275 13:-0.806268 14:0.213675 15:0.481653
16:-0.221505 17:0.151303 18:-0.549101 19:-0.765079 20:0.245536 21:0.231962 22:-0.843555
-1.607719553857081 1:-0.177211 2:0.475751 3:0.030303 4:-0.889868 5:-0.983193 6:-0.208494
7:0.419713 8:0.241586 9:0.592734 10:-0.628499 11:-0.615044 12:0.644034 13:-0.975596 14:-0.965812
15:-0.402086 16:0.37558 17:0.521479 18:0.169887 19:-0.746032 20:-0.580357 21:0.723885 22:-0.967162
-1.028816981696234 1:0.786089 2:0.264434 3:0.0909091 4:-0.921601 5:-0.94958 6:0.106917 7:0.192512
8:0.148701 9:0.221406 10:-1 11:-0.49115 12:0.911668 13:-0.971137 14:-0.931624 15:0.153599
16:0.644822 17:0.709556 18:0.443183 19:-0.619048 20:-0.473214 21:0.916864 22:-0.960438
-3.12725721332852 1:-0.444949 2:-0.159353 3:-1 4:-0.486259 5:0.563025 6:-0.426797 7:-0.123519 8:-0.0625467
9:-0.0625446 10:-0.587786 11:0.619469 12:0.300813 13:-0.701599 14:0.418803 15:-0.342861
16:-0.398223 17:-0.181577 18:-0.37858 19:-0.555556 20:0.589286 21:0.380142 22:-0.783554
-3.6 1:0.161832 2:-0.43418 3:-1 4:-0.524422 5:0.428571 6:-0.748168 7:-0.293489 8:-0.202224 9:-0.202654
10:-0.526718 11:0.146018 12:-0.00505383 13:-0.484929 14:0.435897 15:-0.812025 16:-0.116578
17:0.203562 18:-0.318956 19:-0.409524 20:0.116071 21:-0.0372604 22:-0.534861
-2.539675142648088 1:0.0800419 2:-0.198614 3:-1 4:-0.600954 5:0.159664 6:-0.531355 7:-0.136106
8:-0.0946087 9:-0.189408 10:-0.552163 11:0.274336 12:0.305208 13:-0.710114 14:0.162393 15:-0.465688
16:-0.00642624 17:0.281902 18:-0.208854 19:-0.555556 20:0.245536 21:0.273745 22:-0.756411
-0.9856015024828235 1:-0.117092 2:-0.110855 3:0.111111 4:-0.74427 5:-0.344538 6:-0.674655 7:-0.0757874
8:-0.0518811 9:-0.12092 10:-0.73028 11:0.20354 12:0.836519 13:-0.698892 14:-0.401709
15:-0.535986 16:0.278196 17:0.546976 18:-0.0399796 19:-0.67619 20:0.0892857 21:0.770838 22:-0.800895
-0.05862156536600584 1:0.374345 2:-0.315242 3:-1 4:0.782433 5:0.647059 6:-0.138209 7:-0.394745
8:-0.403635 9:-0.281301 10:-0.521628 11:0.5 12:0.152714 13:-0.720312 14:0.675214 15:-0.172139
16:-0.668926 17:-0.582532 18:-0.428717 19:-0.568254 20:0.540179 21:0.164333 22:-0.764479
-3.6 1:-0.211465 2:0.0150115 3:-0.996768 4:-0.443742 5:-0.0588235 6:-0.464796 7:-0.0218951 8:-0.0178194
9:0.00555184 10:-0.709924 11:-0.455752 12:-0.317073 13:-0.693324 14:-0.025641 15:-0.385606
16:-0.151498 17:0.0853315 18:-0.302521 19:-0.701587 20:-0.450893 21:-0.327159 22:-0.759434
-0.2977339018112915 1:0.241524 2:-0.41224 3:-1 4:-0.0480141 5:0.848739 6:-0.0594809 7:-0.0988589
8:0.0326147 9:-0.220599 10:-0.755725 11:0.283186 12:-0.113601 13:-0.694347 14:0.897436
15:0.0446762 16:-0.183341 17:0.342069 18:-0.678024 19:-0.580952 20:0.299107 21:-0.122981 22:-0.750541
-0.1656419629395251 1:0.878364 2:-1 3:-1 4:-0.0627398 5:0.815126 6:-0.253446 7:-0.489383 8:-0.301206
9:-0.597457 10:-0.791349 11:0.247788 12:-0.125027 13:-0.675517 14:0.82906 15:-0.219003
16:-0.518798 17:-0.108693 18:-0.781897 19:-0.707937 20:0.28125 21:-0.107043 22:-0.742114
-0.3018456460836597 1:0.986718 2:-0.93418 3:-1 4:0.969719 5:0.764706 6:0.013287 7:-0.619137 8:-0.37941
9:-0.571932 10:-0.592875 11:0.216814 12:-0.123709 13:-0.694534 14:0.709402 15:0.00965624
16:-0.297623 17:0.0465646 18:-0.530053 19:-0.396825 20:0.205357 21:-0.105751 22:-0.760519
-2.668626145209309 1:-0.41489 2:0.162818 3:-1 4:0.825158 5:-0.226891 6:0.400224 7:0.152722
8:0.295857 9:0.0273676 10:-0.659033 11:-0.349558 12:-0.0670182 13:-0.8506 14:-0.264957
15:0.438651 16:-0.228753 17:0.0285967 18:-0.411014 19:-0.479365 20:-0.352679 21:-0.0372604 22:-0.879883
-1.2372144351773 1:-0.375743 2:0.264434 3:-1 4:0.875143 5:-0.193277 6:0.320502 7:0.198571
8:0.0840406 9:0.370015 10:-0.577608 11:-0.230088 12:0.0591079 13:-0.865013 14:-0.196581
15:0.349813 16:0.0593672 17:0.34275 18:-0.171023 19:-0.663492 20:-0.196429 21:0.105751 22:-0.884692
-2.076531192233184 1:-0.375743 2:0.0969977 3:-0.992323 4:0.38152 5:-0.159664 6:-0.225133 7:-0.0121166
8:0.0794394 9:-0.142781 10:-0.587786 11:0.119469 12:0.457702 13:-0.799064 14:-0.247863
15:-0.182439 16:-0.115679 17:-0.0241839 18:-0.148075 19:-0.479365 20:-0.0491071 21:0.348266 22:-0.841525
-0.1213928415272433 1:-0.51835 2:0.110855 3:-1 4:0.981748 5:0.394958 6:-0.194338 7:-0.251551 8:-0.0617826
9:-0.449308 10:-0.776081 11:0.672566 12:0.477477 13:-0.740634 14:0.350427 15:-0.190936
16:-0.431963 17:-0.205175 18:-0.568244 19:-0.580952 20:0.598214 21:0.445186 22:-0.786837
-1.890170965438599 1:-0.793778 2:0.140878 3:-0.999998 4:0.908949 5:0.327731 6:-0.134981 7:-0.293415
8:-0.2085 9:-0.26172 10:-0.709924 11:0.513274 12:0.390903 13:-0.729908 14:0.316239 15:-0.257371
16:-0.185055 17:0.070353 18:-0.393481 19:-0.631746 20:0.459821 21:0.349558 22:-0.786517

-0.9013317321267164 1:-0.3331 2:-0.147806 3:-0.999992 4:0.981748 5:0.394958 6:-0.111635
7:0.0788313 8:0.266982 9:-0.172844 10:-0.709924 11:0.575221 12:0.388706 13:-0.743767 14:0.264957
15:-0.202008 16:-0.0499982 17:0.352524 18:-0.415028 19:-0.568254 20:0.535714 21:0.462847 22:-
0.782027
-2.400061946896066 1:-0.0241174 2:-0.180139 3:-1 4:0.945453 5:0.361345 6:0.00484292 7:-0.221892
8:-0.268153 9:0.0298556 10:-0.679389 11:0.588496 12:0.428697 13:-0.750795 14:0.299145 15:-
0.287756 16:-0.26283 17:0.032978 18:-0.34323 19:-0.51746 20:0.508929 21:0.406418 22:-0.801959
-2.992228276076969 1:-0.580566 2:-0.0762125 3:-0.999996 4:0.981748 5:0.394958 6:-0.131007 7:-
0.127401 8:-0.0109583 9:-0.21567 10:-0.720102 11:0.535398 12:0.353988 13:-0.731043 14:0.316239
15:-0.227243 16:0.0724585 17:0.373576 18:-0.343854 19:-0.428571 20:0.575893 21:0.456386 22:-
0.7751
-1.559908856079281 1:-0.314226 2:-0.0496536 3:-1 4:0.69076 5:0.12605 6:-0.0920154 7:-0.262153 8:-
0.290508 9:-0.0701698 10:-0.684478 11:0.212389 12:0.274445 13:-0.758271 14:0.0769231 15:-0.285181
16:0.13592 17:0.549427 18:-0.298602 19:-0.695238 20:0.232143 21:0.335344 22:-0.79496
-0.9221418989660528 1:-0.38483 2:-0.0381062 3:-0.999994 4:0.854402 5:0.277311 6:-0.144915 7:-
0.266212 8:-0.199607 9:-0.211062 10:-0.709924 11:0.517699 12:0.437047 13:-0.731155 14:0.145299
15:-0.183984 16:-0.177062 17:0.217576 18:-0.373998 19:-0.51746 20:0.392857 21:0.43571 22:-
0.793852
-2.488587258405685 1:-0.357567 2:0.366051 3:-0.79798 4:-0.608006 5:-0.831933 6:-0.62871
7:0.455857 8:0.319367 9:0.619871 10:-0.6743 11:-0.535398 12:0.44232 13:-0.815763 14:-0.811966
15:-0.770568 16:0.548225 17:0.617819 18:0.383389 19:-0.701587 20:-0.526786 21:0.426233 22:-
0.797122
-2.946587622272495 1:-0.428172 2:0.213626 3:-0.999998 4:-0.153583 5:-0.176471 6:-0.455358 7:-
0.246081 8:-0.193721 9:-0.200949 10:-0.669211 11:0.216814 12:0.601406 13:-0.776249 14:-0.145299
15:-0.560706 16:0.0469642 17:0.231763 18:-0.137926 19:-0.580952 20:0.370536 21:0.748869 22:-
0.792508
-3.6 1:-0.124782 2:0.0692841 3:-0.999677 4:-0.025407 5:0.00840336 6:0.044828 7:-0.197527 8:-
0.0886092 9:-0.310995 10:-0.506361 11:0.225664 12:0.397056 13:-0.771096 14:-0.00854701 15:-
0.158749 16:-0.0252765 17:0.368278 18:-0.485312 19:-0.555556 20:0.191964 21:0.378419 22:-0.796113
-3.6 1:-0.407899 2:0.132794 3:-1 4:-0.173701 5:-0.00840336 6:-0.525643 7:0.0336588 8:0.224872 9:-
0.196779 10:-0.643766 11:0.0840708 12:0.259064 13:-0.677451 14:-0.0769231 15:-0.523111 16:-
0.110805 17:0.161126 18:-0.349876 19:-0.612698 20:-0.00892857 21:0.218178 22:-0.752822
-1.626125961000681 1:-0.648375 2:0.235566 3:-0.999996 4:-0.468837 5:0.462185 6:-0.404694 7:-
0.229406 8:-0.145335 9:-0.198 10:-0.709924 11:0.415929 12:0.202373 13:-0.685113 14:0.230769 15:-
0.176516 16:-0.593033 17:-0.51917 18:-0.302786 19:-0.701587 20:0.236607 21:0.208271 22:-0.788352
-2.068722239851308 1:-0.403006 2:0.502309 3:-0.191919 4:-0.96723 5:-0.983193 6:-0.856948
7:0.720686 8:0.647612 9:0.723306 10:-0.531807 11:-1 12:-0.582509 13:-0.711787 14:-0.965812 15:-
0.85065 16:0.672339 17:0.965377 18:0.193858 19:0.701587 20:-1 21:-0.586905 22:-0.734577
-0.009736910740550647 1:0.540021 2:-0.277136 3:-1 4:1 5:-0.310924 6:-0.531603 7:-0.187645 8:-
0.118623 9:-0.161902 10:-0.56743 11:-0.349558 12:0.00988794 13:-0.698605 14:-0.350427 15:-
0.573581 16:-0.146034 17:0.175693 18:-0.335864 19:-0.542857 20:-0.299107 21:0.125135 22:-0.764244
-2.203116017217018 1:0.300245 2:-0.125866 3:-1 4:0.970756 5:-0.327731 6:-0.639389 7:-0.122327
8:0.0373947 9:-0.237209 10:-0.597964 11:-0.402655 12:-0.0490002 13:-0.682673 14:-0.333333 15:-
0.619415 16:-0.159659 17:0.280495 18:-0.599655 19:-0.619048 20:-0.383929 21:-0.0135688 22:-
0.755087
-0.9283467294476343 1:-0.535128 2:0.145497 3:-0.79798 4:0.847558 5:-0.445378 6:-0.177201 7:-
0.326633 8:-0.249423 9:-0.406459 10:-0.857506 11:-0.513274 12:-0.101296 13:-0.873623 14:-0.418803
15:-0.201493 16:-0.0939495 17:0.123039 18:-0.249676 19:-0.777778 20:-0.522321 21:-0.136765 22:-
0.901355
-0.7867629932876574 1:-0.535128 2:0.095843 3:-0.373737 4:0.847558 5:-0.445378 6:0.108407 7:-
0.133076 8:0.00239001 9:-0.220093 10:-0.78626 11:-0.544248 12:-0.154032 13:-0.890535 14:-0.452991
15:0.0992661 16:-0.124149 17:0.158263 18:-0.31445 19:-0.777778 20:-0.566964 21:-0.172087 22:-
0.897458
-0.4654094018185496 1:0.691017 2:-0.211316 3:-1 4:0.911646 5:-0.747899 6:-0.628958 7:0.549084
8:0.360095 9:0.762261 10:-0.689567 11:-0.668142 12:-0.0239508 13:-0.792292 14:-0.811966 15:-
0.641818 16:0.622797 17:0.735195 18:0.437882 19:-0.377778 20:-0.709821 21:-0.0165841 22:-0.842454
-0.7108599574961633 1:0.972038 2:-0.381062 3:-1 4:0.911646 5:-0.747899 6:-0.603378 7:0.44549
8:0.394368 9:0.463498 10:-0.659033 11:-0.734513 12:-0.175126 13:-0.797311 14:-0.794872 15:-
0.631775 16:0.411539 17:0.509553 18:0.368696 19:-0.263492 20:-0.700893 21:-0.0191686 22:-0.832141
-0.5336266171710357 1:0.907725 2:-0.181293 3:-1 4:0.955823 5:-0.731092 6:-0.660002 7:0.490074
8:0.50517 9:0.405722 10:-0.628499 11:-0.646018 12:-0.00637223 13:-0.784854 14:-0.777778 15:-
0.62096 16:0.713762 17:0.827787 18:0.442009 19:-0.320635 20:-0.705357 21:-0.0570752 22:-0.82507
-3.6 1:-0.532331 2:-0.0034642 3:-0.953535 4:-0.273255 5:0.546218 6:-0.514467 7:-0.201585 8:-
0.297743 9:-0.100302 10:-0.613232 11:0.0221239 12:-0.171611 13:-0.699478 14:0.504274 15:-0.520536
16:-0.101661 17:-0.0453309 18:-0.0233364 19:-0.504762 21:-0.171226 22:-0.758138
-3.6 1:-0.411395 2:0.475751 3:-0.69697 4:-0.732656 5:-0.630252 6:-0.653297 7:0.35816 8:0.309222
9:0.403303 10:-0.180662 11:-0.610619 12:-0.0678972 13:-0.816243 14:-0.606838 15:-0.599846
16:0.379345 17:0.403723 18:0.259067 19:-0.542857 20:-0.59375 21:-0.0613827 22:-0.853904
-0.160177027597109 1:0.641384 2:0.163972 3:-0.884848 4:-0.79301 5:-0.882353 6:-0.324972
7:0.162765 8:0.105323 9:0.287498 10:-0.832061 11:-0.486726 12:0.70512 13:-0.912158 14:-0.897436
15:-0.343118 16:0.162145 17:0.290997 18:0.0349431 19:-0.638095 20:-0.473214 21:0.795822 22:-
0.930831
-0.4565607028875712 1:0.955959 2:-0.97806 3:-1 4:1 5:0.445378 6:0.186887 7:-0.624121 8:-0.289809
9:-0.813817 10:-0.842239 11:0.0353982 12:-0.111404 13:-0.749634 14:0.487179 15:0.285181 16:-
0.246564 17:0.106004 18:-0.532685 19:-0.663492 20:0.0357143 21:-0.13418 22:-0.790657

-3.6 1:-0.403006 2:0.248268 3:-1 4:0.102147 5:0.0588235 6:0.114864 7:-0.0035291 8:0.119354 9:-0.0293718 10:-0.603053 11:0.0132743 12:0.120633 13:-0.807712 14:0.042735 15:-0.128364 16:-0.0243073 17:0.236002 18:-0.376383 19:-0.67619 20:-0.0669643 21:0.0514753 22:-0.850518
-0.4337488314378258 1:-0.736456 2:0.308314 3:-0.874747 4:-0.973867 5:-0.697479 6:-0.0254564 7:0.0919919 8:0.209882 9:-0.103043 10:-0.770992 11:-0.477876 12:0.300813 13:-0.88347 14:-0.675214 15:0.0395262 16:0.340085 17:0.483171 18:0.231101 19:-0.574603 20:-0.477679 21:0.263407 22:-0.891687
-1.9514203958442 1:0.403006 2:-0.115473 3:-1 4:-0.107954 5:-0.361345 6:0.235316 7:-0.136326 8:-0.243635 9:0.00329425 10:-0.491094 11:-0.455752 12:-0.0977807 13:-0.844366 14:-0.333333 15:0.214626 16:-0.027243 17:0.00635834 18:0.015668 19:-0.384127 20:-0.508929 21:-0.191902 22:-0.872102
-2.968617515406855 1:-0.124782 2:0.30254 3:-0.985455 4:-0.68516 5:-1 6:-0.37812 7:0.263447 8:0.224596 9:0.198922 10:-0.547074 11:-0.59292 12:0.785542 13:-0.925063 14:-0.982906 15:-0.427578 16:0.32219 17:0.426104 18:-0.0211022 19:-0.638095 20:-0.580357 21:0.781607 22:-0.935584
-1.960987134046766 1:0.0415938 2:0.566975 3:0.878788 4:-0.488541 5:-0.915966 6:-0.813237 7:0.420992 8:0.503268 9:0.296551 10:0.170483 11:-0.99115 12:-0.626895 13:-0.669986 14:-0.897436 15:-0.806875 16:0.561091 17:0.683854 18:0.291085 19:0.225397 20:-0.991071 21:-0.630411 22:-0.736563
-1.887010412254054 1:0.357567 2:0.13164 3:-0.737374 4:-0.911231 5:-0.781513 6:-0.41289 7:0.355322 8:0.374711 9:0.255638 10:-0.765903 11:-0.747788 12:-0.159745 13:-0.85888 14:-0.760684 15:-0.602163 16:0.787351 17:0.760945 18:0.665016 19:-0.492063 20:-0.830357 21:-0.362481 22:-0.860353
-0.8717054952085065 1:-0.491087 2:0.170901 3:-1 4:-0.0635694 5:0.092437 6:-0.357507 7:-0.0126165 8:0.135954 9:-0.201018 10:-0.56743 11:0.274336 12:0.36849 13:-0.733995 14:0.0769231 15:-0.400798 16:0.267451 17:0.43213 18:0.0795899 19:-0.415873 20:0.267857 21:0.376265 22:-0.803387
-0.5909171887386574 1:-0.0597693 2:-0.431871 3:-1 4:0.22576 5:0.781513 6:-0.431516 7:-0.254683 8:-0.193884 9:-0.307978 10:-0.755725 11:0.0663717 12:-0.241925 13:-0.59393 14:0.811966 15:-0.368353 16:-0.214903 17:0.139061 18:-0.520359 19:-0.688889 20:0.0535714 21:-0.262977 22:-0.679206
-0.5863860863205241 1:0.273681 2:0.438799 3:-0.919192 4:-0.91372 5:-0.983193 6:0.157581 7:0.273432 8:0.0867883 9:0.390979 10:-0.765903 11:-0.526549 12:0.928807 13:-0.989034 14:-0.965812 15:0.235741 16:0.624511 17:0.748766 18:0.36987 19:-0.822222 20:-0.491071 21:1 22:-0.974313
-1.694973078926043 1:0.150647 2:-0.264434 3:-1 4:-0.156279 5:-0.210084 6:0.154104 7:-0.144105 8:-0.0559132 9:-0.151259 10:-0.659033 11:-0.159292 12:0.164579 13:-0.842378 14:-0.196581 15:0.0292262 16:0.0368648 17:0.199624 18:-0.162692 19:-0.657143 20:-0.138393 21:0.179841 22:-0.871181
-0.3741974414838726 1:0.819643 2:-0.00923788 3:-0.965657 4:-0.939853 5:-0.798319 6:0.198063 7:0.134165 8:0.19616 9:0.14239 10:-0.928753 11:-0.438053 12:0.597451 13:-0.91382 14:-0.846154 15:0.0338612 16:0.307174 17:0.459825 18:0.127323 19:-0.75873 20:-0.455357 21:0.691579 22:-0.916202
-2.273040088367057 1:0.186298 2:-0.116628 3:-0.999879 4:-0.316188 5:-0.226891 6:-0.493108 7:-0.204982 8:-0.141011 9:-0.120712 10:-0.470738 11:0.0442478 12:0.445836 13:-0.742291 14:-0.282051 15:-0.463886 16:-0.0867156 17:0.208022 18:-0.351712 19:-0.695238 20:0.0446429 21:0.52143 22:-0.806956
-2.501826994570088 1:-0.63719 2:0.196305 3:-0.858586 4:-0.296277 5:-0.193277 6:-0.540047 7:-0.0390554 8:0.0998439 9:-0.205464 10:-0.394402 11:0.0486726 12:0.408482 13:-0.741902 14:-0.247863 15:-0.382516 16:-0.100945 17:0.260376 18:-0.55232 19:-0.466667 20:0.0758929 21:0.514969 22:-0.79892
-0.5294492863270115 1:-0.407899 2:-0.0831409 3:-0.999939 4:-0.156694 5:0.0420168 6:-0.380604 7:-0.456474 8:-0.273827 9:-0.426133 10:-0.684478 11:0.225664 12:0.365414 13:-0.699867 14:-0.0769231 15:-0.446376 16:-0.210001 17:-0.12209 18:-0.28904 19:-0.492063 20:0.15625 21:0.409864 22:-0.787395
-3.6 1:-0.180007 2:0.144342 3:-0.999838 4:-0.326143 5:-0.243697 6:-0.460574 7:0.0642884 8:0.115387 9:-0.00889217 10:-0.379135 11:0.0707965 12:0.503406 13:-0.761074 14:-0.282051 15:-0.471096 16:0.155332 17:0.49013 18:-0.330998 19:-0.498413 20:0.0625 21:0.541676 22:-0.803782
-1.259468460313572 1:-0.313527 2:0.095843 3:0.171717 4:-0.588302 5:-0.445378 6:-0.423569 7:-0.0301885 8:0.210288 9:-0.34509 10:-0.806616 11:0.0309735 12:0.755658 13:-0.789591 14:-0.470085 15:-0.533926 16:0.078962 17:0.354533 18:-0.1956 19:-0.619048 20:0.0669643 21:0.856989 22:-0.825357
-0.3877509456706142 1:0.621111 2:0.0323326 3:-1 4:0.821217 5:-0.613445 6:0.379113 7:-0.0636414 8:-0.0419309 9:-0.0901887 10:-0.867684 11:-0.207965 12:0.666007 13:-0.941543 14:-0.641026 15:0.467233 16:-0.0733153 17:0.200177 18:-0.269784 19:-0.834921 20:-0.236607 21:0.672625 22:-0.949774
-0.1477188315336815 1:0.354771 2:-0.0635104 3:-1 4:0.821217 5:-0.613445 6:0.445424 7:-0.042658 8:0.0404351 9:-0.221129 10:-0.877863 11:-0.265487 12:0.55746 13:-0.934755 14:-0.623932 15:-0.113944 16:0.0437616 17:0.245287 18:-0.119863 19:-0.809524 20:-0.299107 21:0.520999 22:-0.928286
-0.5867796549126889 1:0.4813 2:0.0311778 3:-1 4:0.821217 5:-0.613445 6:0.377872 7:-0.0658912 8:0.055019 9:-0.201686 10:-0.872774 11:-0.19469 12:0.687541 13:-0.937936 14:-0.641026 15:0.458736 16:0.0751835 17:0.338226 18:-0.322781 19:-0.834921 20:-0.227679 21:0.695025 22:-0.947644
-3.6 1:-0.667948 2:0.255196 3:-1 4:0.847765 5:0.193277 6:-0.53061 7:0.179734 8:0.231116 9:0.162616 10:-0.521628 11:0.29646 12:0.300374 13:-0.690458 14:-0.00854701 15:-0.417278 16:0.164069 17:0.31643 18:-0.0405476 19:-0.587302 20:0.214286 21:0.403834 22:-0.797469
-2.418826937551479 1:-0.505068 2:0.124711 3:-1 4:0.923883 5:0.260504 6:-0.185148 7:0.00438197 8:0.00681234 9:0.0507498 10:-0.557252 11:0.274336 12:0.223907 13:-0.721825 14:0.128205 15:-0.327411 16:-0.677916 17:-0.297782 18:-0.849019 19:-0.606349 20:0.1875 21:0.245746 22:-0.795742
-0.6084225507020989 1:0.0646627 2:-0.267898 3:-1 4:0.885928 5:0.226891 6:-0.382342 7:-0.138694 8:-0.0499301 9:-0.174802 10:-0.618321 11:0.384956 12:0.357064 13:-0.691806 14:0.0769231 15:-0.363203 16:0.125076 17:0.325952 18:-0.00866239 19:-0.269841 20:0.290179 21:0.398234 22:-0.772998

-0.5930346910637662 1:-0.189794 2:-0.0773672 3:-1 4:0.935705 5:0.596639 6:-0.154352 7:-0.0155869
8:0.212386 9:-0.286139 10:-0.73028 11:0.0663717 12:-0.161064 13:-0.669528 14:0.623932 15:-
0.109051 16:-0.122534 17:0.324102 18:-0.482907 19:-0.587302 20:0.111607 21:-0.140211 22:-0.744706
-1.75110785901865 1:-0.444949 2:0.300231 3:-0.992323 4:-0.659235 5:-0.428571 6:0.531355 7:0.15481
8:-0.0304848 9:0.432122 10:-0.501272 11:-0.0530973 12:0.592617 13:-0.888031 14:-0.401709 15:-
0.0802111 16:0.134909 17:0.301167 18:0.026044 19:-0.733333 20:-0.0267857 21:0.592505 22:-0.89144
-3.6 1:-0.189095 2:0.100462 3:-1 4:-0.276781 5:0.563025 6:-0.000620887 7:-0.290651 8:-0.140799
9:-0.288005 10:-0.755725 11:-0.0884956 12:-0.269172 13:-0.745222 14:0.487179 15:-0.104674 16:-
0.267648 17:-0.0885581 18:-0.452423 19:-0.822222 20:-0.0982143 21:-0.245316 22:-0.802235
-3.6 1:-0.377141 2:0.113164 3:-0.89697 4:-0.639739 5:-0.310924 6:-0.463306 7:0.346146 8:0.272266
9:0.434956 10:-0.389313 11:-0.389381 12:-0.0472424 13:-0.791887 14:-0.316239 15:-0.443028
16:0.453229 17:0.588479 18:0.255167 19:-0.415873 20:-0.450893 21:-0.125996 22:-0.811431
-3.6 1:-0.377141 2:0.125866 3:-0.777778 4:-0.639739 5:-0.310924 6:-0.468024 7:0.351998 8:0.323952
9:0.372826 10:-0.557252 11:-0.415929 12:-0.0872336 13:-0.789857 14:-0.316239 15:-0.445861
16:0.359104 17:0.538973 18:0.0628899 19:-0.377778 20:-0.459821 21:-0.138488 22:-0.814083
-1.53253038641823 1:-0.590353 2:0.533487 3:-0.414141 4:-0.679975 5:-0.798319 6:-0.165777
7:0.150575 8:0.276688 9:-0.0194199 10:-0.770992 11:-0.606195 12:0.195342 13:-0.875131 14:-
0.777778 15:-0.149736 16:0.38169 17:0.609674 18:0.0940745 19:-0.485714 20:-0.598214 21:0.185871
22:-0.892649
-0.05927329405565909 1:0.984621 2:-0.906467 3:-1 4:-0.054651 5:0.932773 6:-0.21818 7:-0.364541
8:-0.261422 9:-0.460872 10:-0.801527 11:0.482301 12:-0.0112063 13:-0.635789 14:0.641026 15:-
0.113171 16:-0.215465 17:-0.289748 18:-0.0193791 19:-0.574603 20:0.450893 21:0.115658 22:-
0.739582
-0.27414113208681 1:-0.142258 2:-0.212471 3:-0.999818 4:-0.500985 5:0.0420168 6:-0.147895
7:0.12243 8:0.343853 9:-0.144855 10:-0.628499 11:0.252212 12:0.395298 13:-0.78593 14:-0.162393
15:0.00785374 16:-0.468694 17:-0.354138 18:-0.399313 19:-0.574603 20:0.129464 21:0.474908 22:-
0.845465
-3.084103827739824 1:0.897938 2:-0.751732 3:-1 4:-0.0938505 5:0.378151 6:-0.271824 7:0.174705
8:0.185982 9:0.117234 10:-0.577608 11:-0.0132743 12:-0.117996 13:-0.710694 14:0.247863 15:-
0.474186 16:0.165713 17:0.343699 18:-0.0103097 19:-0.631746 20:-0.290179 21:-0.301745 22:-
0.742896
-0.07509700426312073 1:0.0569731 2:-0.30254 3:-1 4:0.89775 5:0.764706 6:-0.283497 7:-0.480222 8:-
0.486522 9:-0.308047 10:-0.73028 11:0.0840708 12:-0.220391 13:-0.696356 14:0.811966 15:-0.235741
16:-0.313411 17:-0.154609 18:-0.325129 19:-0.75873 20:0.169643 21:-0.178548 22:-0.743446
-3.6 1:-0.564488 2:0.191686 3:-1 4:0.89775 5:0.764706 6:-0.22414 7:-0.387245 8:-0.28594 9:-
0.384851 10:-0.801527 11:0.150442 12:-0.173368 13:-0.703102 14:0.794872 15:-0.310416 16:-0.132675
17:0.0504081 18:-0.297314 19:-0.75873 20:0.147321 21:-0.188886 22:-0.743127
-0.6021023551917601 1:0.499476 2:0.127021 3:-1 4:-0.907083 5:-0.983193 6:-0.791879 7:0.528189
8:0.515413 9:0.470432 10:-0.938931 11:-0.646018 12:0.546473 13:-0.742387 14:-0.965812 15:-
0.575126 16:0.871489 17:0.7928 18:0.777107 19:-0.771429 20:-0.642857 21:0.524876 22:-0.804392
-3.6 1:-0.357567 2:0.523095 3:-0.981414 4:-0.842995 5:-0.747899 6:-0.00633304 7:0.298768
8:0.290199 9:0.305213 10:-0.664122 11:-0.469027 12:0.411997 13:-0.912866 14:-0.726496 15:-
0.0498262 16:0.133504 17:0.283073 18:-0.010783 19:-0.587302 20:-0.477679 21:0.360758 22:-0.927847
-0.6816476664344064 1:-0.257602 2:0.498845 3:-0.975758 4:-0.858343 5:-0.798319 6:-0.404694
7:0.276976 8:0.266608 9:0.283973 10:-0.648855 11:-0.473451 12:0.513953 13:-0.888702 14:-0.777778
15:-0.369126 16:0.323623 17:0.336502 18:0.267455 19:-0.574603 20:-0.5 21:0.415895 22:-0.898359
-3.513120433434259 1:-0.335897 2:0.242494 3:-0.90303 4:-0.745515 5:-0.428571 6:0.100459
7:0.273667 8:0.263421 9:0.247506 10:-0.603053 11:-0.0575221 12:0.585586 13:-0.881435 14:-0.401709
15:-0.0585812 16:0.118882 17:0.169524 18:0.0971798 19:-0.771429 20:-0.0803571 21:0.514969 22:-
0.89363
-1.129029374297946 1:-0.00803915 2:0.184758 3:-0.717172 4:-0.766048 5:-0.495798 6:0.394015
7:0.248978 8:0.102364 9:0.450068 10:-0.740458 11:-0.172566 12:0.513514 13:-0.909989 14:-0.470085
15:-0.0155787 16:0.256059 17:0.202518 18:0.292259 19:-0.695238 20:-0.111607 21:0.572259 22:-
0.897307
-0.4308067879496676 1:0.00943726 2:-0.0669746 3:-1 4:0.900239 5:0.159664 6:0.148889 7:-0.264491
8:-0.165317 9:-0.426202 10:-0.750636 11:-0.163717 12:-0.13118 13:-0.785445 14:0.128205
15:0.110081 16:-0.102588 17:0.396527 18:-0.551771 19:-0.593651 20:-0.205357 21:-0.153995 22:-
0.824033
-2.486786956857335 1:0.221251 2:-0.224018 3:-1 4:0.900239 5:0.159664 6:-0.0356389 7:-0.186792 8:-
0.174129 9:-0.193739 10:-0.689567 11:-0.190265 12:-0.160185 13:-0.77327 14:0.111111 15:-0.180636
16:-0.0656038 17:0.142746 18:-0.213114 19:-0.644444 20:-0.169643 21:-0.108766 22:-0.822331
-0.9829078482817289 1:0.660958 2:-0.323326 3:-1 4:0.541222 5:-0.142857 6:-0.280268 7:-0.362365
8:-0.144474 9:-0.467783 10:-0.750636 11:-0.517699 12:-0.343002 13:-0.793869 14:-0.128205 15:-
0.201493 16:-0.120104 17:0.218287 18:-0.351523 19:-0.68254 20:-0.540179 21:-0.374542 22:-0.829557
-1.60212096311091 1:0.0751485 2:0.0508083 3:-0.89899 4:0.187182 5:0.00840336 6:-0.748665
7:0.547922 8:0.383296 9:0.725656 10:-0.419847 11:-0.274336 12:-0.150956 13:-0.594996 14:-
0.0598291 15:-0.76516 16:0.642238 17:0.616696 18:0.567259 19:-0.339683 20:-0.325893 21:-0.16261
22:-0.67501
-0.7356490029844409 1:0.108703 2:0.0207852 3:-0.69697 4:0.187182 5:0.00840336 6:-0.776481
7:0.642929 8:0.396904 9:0.922942 10:-0.155216 11:-0.309735 12:-0.194463 13:-0.537823 14:0.042735
15:-0.755633 16:0.614285 17:0.552939 18:0.591059 19:-0.352381 20:-0.299107 21:-0.201378 22:-
0.658966
-0.6272705394738248 1:0.25201 2:-0.397229 3:-0.999657 4:-0.0716582 5:0.193277 6:-0.0423445 7:-
0.141046 8:-0.0485969 9:-0.205142 10:-0.582697 11:0.327434 12:0.330697 13:-0.75663 14:0.0940171
15:-0.169306 16:-0.222699 17:0.0200557 18:-0.419856 19:-0.44127 20:0.25 21:0.33879 22:-0.808811

-0.05443265340706308 1:-0.306536 2:-0.051963 3:-1 4:-0.260396 5:-0.12605 6:-0.556439 7:0.0455107
8:0.144279 9:-0.0111728 10:-0.745547 11:0.0486726 12:0.337289 13:-0.731139 14:-0.0940171 15:-
0.695378 16:0.101844 17:0.36184 18:-0.136544 19:-0.580952 20:0.0401786 21:0.292699 22:-0.758086
-0.5524419882742706 1:0.117092 2:0.100462 3:-0.999999 4:-0.332573 5:-0.428571 6:-0.458835 7:-
0.193483 8:-0.378353 9:0.126264 10:-0.638677 11:-0.393805 12:0.0630631 13:-0.822242 14:-0.57265
15:-0.125016 16:0.270625 17:0.424111 18:0.0513969 19:-0.638095 20:-0.491071 21:0.0889511 22:-
0.911265
-0.7829503586831148 1:0.319818 2:-0.466513 3:-1 4:0.595562 5:0.647059 6:-0.0276915 7:-0.544775
8:-0.274233 9:-0.640144 10:-0.791349 11:1 12:0.539881 13:-0.695402 14:0.675214 15:-0.0153212 16:-
0.175629 17:0.106526 18:-0.399445 19:-0.701587 20:1 21:0.511523 22:-0.754952
-0.2073493818364658 1:-0.268088 2:0.127021 3:-0.985253 4:-0.697604 5:-0.613445 6:-0.606855
7:0.106064 8:0.121387 9:0.110945 10:-0.725191 11:-0.539823 12:0.0437267 13:-0.75046 14:-0.726496
15:-0.403116 16:0.50124 17:0.583354 18:0.317498 19:-0.75873 20:-0.665179 21:-0.0372604 22:-
0.83585
-3.6 1:-0.357567 2:-0.0288684 3:-1 4:0.736596 5:0.344538 6:-0.563641 7:-0.290019 8:-0.0869183 9:-
0.214449 10:-0.597964 11:0.150442 12:0.0498791 13:-0.666192 14:0.333333 15:-0.463371 16:-0.192598
17:-0.00382766 18:-0.316267 19:-0.663492 20:0.138393 21:0.0437217 22:-0.749317
-3.6 1:0.0541769 2:-0.26097 3:-1 4:0.769574 5:0.378151 6:-0.411648 7:-0.00714643 8:-0.0176243
9:0.0216084 10:-0.613232 11:0.212389 12:0.0828389 13:-0.696068 14:0.350427 15:-0.401313 16:-
0.0826421 17:0.0805865 18:-0.173882 19:-0.453968 20:0.138393 21:0.0333836 22:-0.764611
-1.994182250979287 1:0.419084 2:0.106236 3:-1 4:-0.269522 5:-0.747899 6:0.559667 7:0.107035
8:0.134458 9:0.0921468 10:-0.898219 11:-0.553097 12:0.224346 13:-0.938272 14:-0.74359 15:0.508433
16:0.276342 17:0.613738 18:-0.117155 19:-0.84127 20:-0.558036 21:0.219039 22:-0.948135
-3.346410148691686 1:-0.191891 2:0.245958 3:-0.777778 4:-0.388779 5:-0.226891 6:0.106668
7:0.252331 8:0.395083 9:0.035131 10:-0.704835 11:-0.353982 12:-0.0766864 13:-0.833933 14:-
0.264957 15:-0.229303 16:-0.460477 17:-0.317032 18:-0.446042 19:-0.860317 20:-0.330357 21:-
0.00409218 22:-0.840975
-2.236276103063065 1:-0.621811 2:0.275982 3:-0.999475 4:-0.346676 5:-0.260504 6:-0.539054
7:0.0289386 8:0.162911 9:-0.0869635 10:-0.618321 11:-0.00884956 12:0.418589 13:-0.757056 14:-
0.230769 15:-0.422171 16:-0.288787 17:0.00657978 18:-0.415009 19:-0.75873 21:0.396511 22:-
0.787814
-0.4821050182091517 1:0.150647 2:-0.341801 3:-1 4:-0.317225 5:-0.210084 6:-0.488141 7:0.122945
8:0.114249 9:0.188601 10:-0.689567 11:-0.0353982 12:0.322786 13:-0.759257 14:-0.247863 15:-
0.332046 16:-0.2092 17:-0.012859 18:-0.315169 19:-0.619048 20:-0.0446429 21:0.358604 22:-0.804839
-0.6349860529356554 1:0.852499 2:-0.39261 3:-1 4:-0.415535 5:-0.378151 6:-0.561406 7:-0.165574
8:-0.0479303 9:-0.2731 10:-0.826972 11:0.216814 12:0.913426 13:-0.745078 14:-0.350427 15:-
0.632033 16:0.324915 17:0.616396 18:-0.0133959 19:-0.542857 20:0.227679 21:0.885419 22:-0.773963
-0.7899119071708065 1:-0.338693 2:0.124711 3:-0.999999 4:-0.346676 5:-0.260504 6:-0.514715 7:-
0.0227039 8:0.0296069 9:-0.0368126 10:-0.715013 11:-0.0619469 12:0.348715 13:-0.753998 14:-
0.264957 15:-0.418823 16:0.118657 17:0.315466 18:-0.0353975 19:-0.638095 20:-0.00446429
21:0.428818 22:-0.797748
-0.0319693985922057 1:0.241524 2:-0.214781 3:-0.973737 4:-0.317225 5:-0.210084 6:-0.535825 7:-
0.0823457 8:-0.0923812 9:-0.0338409 10:-0.638677 12:0.369809 13:-0.751008 14:-0.247863 15:-
0.452813 16:0.32941 17:0.570986 18:-0.0403014 19:-0.701587 20:-0.0803571 21:0.311652 22:-0.792005
-0.0509952644165541 1:0.501573 2:-0.460739 3:-0.931313 4:-0.0884579 5:0.109244 6:-0.324475
7:0.119945 8:0.27529 9:-0.107005 10:-0.73028 11:0.146018 12:0.219512 13:-0.755101 14:0.128205
15:-0.454101 16:-0.00156618 17:0.241158 18:-0.287109 19:-0.67619 20:0.09375 21:0.151411 22:-
0.790541
-2.556954523555175 1:0.156938 2:-0.384527 3:-1 4:0.0144146 5:0.277311 6:-0.519682 7:-0.20779 8:-
0.152327 9:-0.227326 10:-0.78626 11:0.362832 12:0.2951 13:-0.69375 14:0.316239 15:-0.736578
16:0.0257541 17:0.447615 18:-0.394238 19:-0.644444 20:0.375 21:0.269869 22:-0.731948
-2.1783002934466939 1:-0.26669 2:-0.0554273 3:-1 4:0.0349476 5:0.310924 6:-0.00856824 7:-0.131474
8:-0.0614574 9:-0.214587 10:-0.776081 11:0.584071 12:0.468249 13:-0.7516 14:0.282051 15:0.0184112
16:-0.307975 17:0.208307 18:-0.68094 19:-0.52381 20:0.459821 21:0.375835 22:-0.802306
-0.05171868507332809 1:-0.114995 2:-0.236721 3:-0.999999 4:-0.0679249 5:0.142857 6:-0.225133 7:-
0.108255 8:-0.0436868 9:-0.154461 10:-0.811705 11:0.252212 12:0.299934 13:-0.75242 14:0.162393
15:-0.228016 16:-0.0761105 17:0.204574 18:-0.26168 19:-0.536508 20:0.308036 21:0.336636 22:-
0.78422
-0.3114228432754523 1:0.932891 2:-0.772517 3:-1 4:0.0349476 5:0.310924 6:-0.107165 7:-0.248728
8:-0.30862 9:-0.151789 10:-0.852417 11:0.473451 12:0.369809 13:-0.757056 14:0.247863 15:-0.129909
16:-0.371029 17:-0.110797 18:-0.521665 19:-0.530159 20:0.455357 21:0.401249 22:-0.805521
-0.7437529561882521 1:0.222649 2:-0.314088 3:-1 4:0.0144146 5:0.277311 6:-0.210729 7:-0.211614
8:-0.17569 9:-0.196181 10:-0.745547 11:0.50885 12:0.429576 13:-0.73432 14:0.316239 15:-0.245268
16:-0.143842 17:0.153122 18:-0.350425 19:-0.593651 20:0.526786 21:0.410295 22:-0.787271
-0.7511483299175563 1:0.979727 2:-0.116628 3:-1 4:0.958519 5:-0.680672 6:0.650565 7:0.0252478 8:-
0.0498 9:-0.218411 10:-0.821883 11:-0.544248 12:0.135135 13:-0.932203 14:-0.709402 15:0.580533
16:0.0309653 17:0.114245 18:-0.183349 19:-0.765079 20:-0.544643 21:0.18501 22:-0.952338
-3.6 1:-0.846907 2:0.379908 3:-1 4:0.0164886 5:0.378151 6:-0.268099 7:0.152737 8:0.286427
9:0.00449216 10:-0.770992 11:0.079646 12:-0.0340584 13:-0.719268 14:0.213675 15:0.0938586
16:0.158563 17:0.300282 18:-0.0726789 19:-0.504762 20:0.0133929 21:0.00839974 22:-0.801884
-2.851977656244539 1:0.362461 2:0.230947 3:0.333333 4:-0.928031 5:-0.94958 6:-0.437973
7:0.0895215 8:0.0638637 9:0.117096 10:-0.669211 11:-0.858407 12:-0.186992 13:-0.86716 14:-
0.931624 15:-0.471353 16:0.439365 17:0.654055 18:0.0236583 19:-0.625397 20:-0.808036 21:-
0.0635365 22:-0.885342
-3.6 1:-0.00803915 2:0.170901 3:-0.999475 4:-0.605724 5:-0.680672 6:-0.540544 7:0.201262
8:0.10451 9:0.393075 10:-0.582697 11:-0.853982 12:-0.499011 13:-0.809524 14:-0.675214 15:-
0.615295 16:0.33368 17:0.388523 18:0.264634 19:-0.536508 20:-0.861607 21:-0.502908 22:-0.82322

-1.564984658229848 1:0.574275 2:0.163972 3:-0.656566 4:-0.72996 5:-0.89916 6:-0.153111 7:0.421757
8:0.332878 9:0.52162 10:-0.745547 11:-0.761062 12:-0.00109866 13:-0.919702 14:-0.880342 15:-
0.487833 16:0.579689 17:0.721656 18:0.396984 19:-0.11746 20:-0.839286 21:-0.235408 22:-0.900038
-3.484365495159456 1:-0.780496 2:0.58545 3:-0.272727 4:-0.657368 5:-0.781513 6:-0.00558798
7:0.167132 8:0.0873248 9:0.102329 10:-0.541985 11:-0.650442 12:0.0705339 13:-0.914012 14:-
0.760684 15:0.0184112 16:0.240412 17:0.319009 18:0.0764089 19:-0.746032 20:-0.660714 21:0.0183071
22:-0.926488
-0.004117281490409305 1:-0.182803 2:0.486143 3:-0.454545 4:-0.686405 5:-0.831933 6:0.410406 7:-
0.138105 8:-0.147547 9:-0.183579 10:-0.608142 11:-0.730088 12:-0.0432872 13:-0.943946 14:-
0.811966 15:0.455646 16:0.341223 17:0.431055 18:-0.0480834 19:-0.733333 20:-0.745536 21:-0.101443
22:-0.936514
-1.073261329247968 1:-0.484796 2:0.633949 3:-0.890909 4:-0.754226 5:-0.94958 6:0.0117968
7:0.150943 8:-0.010373 9:0.415536 10:-0.628499 11:-0.89823 12:-0.313557 13:-0.911305 14:-0.931624
15:0.186816 16:0.62666 17:0.677496 18:0.398764 19:-0.180952 20:-0.897321 21:-0.321129 22:-0.92084
-2.138861439070299 1:0.221251 2:-0.401848 3:-1 4:0.445608 5:0.983193 6:-0.193592 7:-0.519292 8:-
0.301532 9:-0.625861 10:-0.73028 11:-0.243363 12:-0.524061 13:-0.60326 14:0.982906 15:-0.240891
16:-0.254795 17:0.071824 18:-0.542531 19:-0.638095 20:-0.191964 21:-0.4874 22:-0.699573
-2.567488626499896 1:-1 2:0.497691 3:-0.996768 4:-0.244426 5:-0.260504 6:-0.0751273 7:-0.126004
8:0.0180145 9:-0.133659 10:-0.857506 11:-0.517699 12:-0.26258 13:-0.804739 14:-0.247863 15:-
0.277713 16:0.21156 17:0.491933 18:-0.128345 19:-0.479365 20:-0.602679 21:-0.383588 22:-0.820253
-1.755400767370126 1:-0.687522 2:0.258661 3:-1 4:-0.266826 5:-0.294118 6:-0.00807153 7:-0.125724
8:-0.0645628 9:-0.0426179 10:-0.832061 11:-0.482301 12:-0.190508 13:-0.815928 14:-0.264957 15:-
0.129651 16:-0.154209 17:-0.130552 18:-0.0591978 19:-0.631746 20:-0.53125 21:-0.275468 22:-
0.847659
-1.304483635625165 1:-0.0632646 2:-0.161663 3:-1 4:-0.655916 5:0.159664 6:0.116851 7:-0.399597
8:-0.368322 9:-0.29669 10:-0.704835 11:-0.123894 12:-0.0947045 13:-0.78414 14:0.128205
15:0.408008 16:-0.304955 17:0.0982222 18:-0.632146 19:-0.606349 20:-0.129464 21:-0.0755977 22:-
0.829393
-3.6 1:-0.0877316 2:-0.0450346 3:-1 4:-0.698019 5:-0.00840336 6:0.207252 7:-0.106241 8:-0.0774233
9:-0.0248105 10:-0.521628 11:-0.252212 12:-0.115359 13:-0.79672 14:0.00854701 15:0.332818 16:-
0.181051 17:0.142604 18:-0.392118 19:-0.669841 20:-0.254464 21:-0.128581 22:-0.830067
-0.1490569347569184 1:0.793778 2:-0.47806 3:-1 4:-0.630613 5:0.109244 6:0.0182541 7:-0.267844 8:-
0.0897636 9:-0.490659 10:-0.73028 11:-0.358407 12:-0.300813 13:-0.767004 14:-0.042735 15:0.277456
16:-0.177624 17:0.11464 18:-0.450624 19:-0.638095 20:-0.46875 21:-0.337067 22:-0.821904
-0.1295018770357079 1:0.900734 2:-0.696305 3:-1 4:-0.579799 5:0.294118 6:0.0284366 7:-0.385981
8:-0.322359 9:-0.214886 10:-0.613232 11:-0.0575221 12:-0.110525 13:-0.753896 14:0.213675 15:-
0.0269087 16:-0.294785 17:0.0499494 18:-0.421466 19:-0.447619 20:-0.151786 21:-0.152703 22:-
0.81366
-0.4878078490410139 1:0.618315 2:-0.192841 3:-1 4:-0.824536 5:-0.596639 6:0.530858 7:-0.13587
8:0.0285826 9:-0.157041 10:-0.735369 11:-0.345133 12:0.379917 13:-0.926374 14:-0.589744
15:0.584138 16:0.165699 17:0.336929 18:0.000710032 19:-0.669841 20:-0.334821 21:0.395219 22:-
0.919253
-0.5026666367258752 1:0.953163 2:-0.555427 3:-1 4:-0.672094 5:-0.0420168 6:-0.189122 7:-0.441196
8:-0.392986 9:-0.334193 10:-0.699746 11:-0.482301 12:-0.356185 13:-0.759859 14:-0.128205 15:-
0.171366 16:-0.0118763 17:0.28214 18:-0.147829 19:-0.555556 20:-0.5 21:-0.326298 22:-0.818885
-0.5504686074028519 1:-0.357567 2:0.371824 3:-0.333333 4:-0.798818 5:-1 6:-0.31777 7:0.384113
8:0.365135 9:0.3171 10:-0.755725 11:-0.827434 12:0.00241705 13:-0.92891 14:-0.982906 15:-0.447406
16:0.575377 17:0.574544 18:0.472739 19:-0.561905 20:-0.883929 21:-0.188886 22:-0.901745
-3.6 1:-0.623209 2:-0.00115473 3:-1 4:0.252515 5:0.882353 6:0.286229 7:-0.235126 8:-0.183998 9:-
0.249764 10:-0.618321 11:0.0265487 12:-0.308284 13:-0.741796 14:0.863248 15:0.135316 16:-0.277494
17:-0.0996932 18:-0.455831 19:-0.650794 20:0.0357143 21:-0.29356 22:-0.78757

Anexo G: Clasificadores optimizados para MLP

Modelo con una capa oculta, 12 neuronas en la capa oculta.

==== Classifier model (full training set) ====

```
Sigmoid Node 0
  Inputs  Weights
  Threshold -0.1143220032977759
  Node 2 -0.8714160713712031
  Node 3 -2.795892976024625
  Node 4 0.8362900790971322
  Node 5 1.3157398064440853
  Node 6 0.26371103393614687
  Node 7 -1.598277878170119
  Node 8 1.9626223286696252
  Node 9 1.7170208468033124
  Node 10 -2.449202958487344
  Node 11 -1.1074549928225603
  Node 12 2.4256677341408026
  Node 13 -3.126721643957588

Sigmoid Node 1
  Inputs  Weights
  Threshold 0.14138351176611463
  Node 2 0.942573890254077
  Node 3 2.7838792051319925
  Node 4 -0.8300039045760115
  Node 5 -1.3710565246360595
  Node 6 -0.2348752983434843
  Node 7 1.5402381056401475
  Node 8 -2.0208178763899003
  Node 9 -1.73811713792009
  Node 10 2.3960702766517823
  Node 11 1.1857388083076863
  Node 12 -2.36326532191992
  Node 13 3.127633545502348

Sigmoid Node 2
  Inputs  Weights
  Threshold 0.17571783245509745
  Attrib ALI_SEQIDE -0.02201895400383992
  Attrib ALI_ZSCO 0.44377054119977877
  Attrib ALI_PBEVAL 0.3659798983208167
  Attrib ALI_TGCOV 0.19513176208732105
  Attrib ML_LEN 0.44257843600133023
  Attrib ML_COMP 0.1670267347187456
  Attrib ML_ZCOMB 0.2735513054584023
  Attrib ML_ZPAIR 0.06071813874170902
  Attrib ML_ZSURF 0.24714280540725664
  Attrib ML_PP 0.6343203267261257
  Attrib ML_COABS -0.1373980980678631
  Attrib ML_COREL 0.0655869111704055
  Attrib ML_RG 0.33356820847622354
  Attrib TF_LEN -0.08110883196912841
  Attrib TF_COMP 0.12965103587509638
  Attrib TF_ZCOMB -0.185967402235069
  Attrib TF_ZPAIR -0.10050624459769843
  Attrib TF_ZSURF -0.14601405820737748
  Attrib TF_PP 0.09455509231225927
  Attrib TF_COABS -0.4684020615072721
  Attrib TF_COREL 0.0799948551049993
  Attrib TF_RG 0.22724802807904784

Sigmoid Node 3
  Inputs  Weights
  Threshold 1.157350735347044
  Attrib ALI_SEQIDE -0.44449150954668465
  Attrib ALI_ZSCO 1.740691687195036
  Attrib ALI_PBEVAL 0.6997832427579412
  Attrib ALI_TGCOV 0.21045382528697076
  Attrib ML_LEN 1.5713917066686156
  Attrib ML_COMP 0.7570115128213367
  Attrib ML_ZCOMB 1.2024373799983907
  Attrib ML_ZPAIR 0.6265951521573284
  Attrib ML_ZSURF 1.2908844760793978
  Attrib ML_PP 2.096560083338782
```

```

Attrib ML_COABS      0.29214094358426057
Attrib ML_COREL     0.7493363558639141
Attrib ML_RG        0.6737390454123325
Attrib TF_LEN       -0.3368795044141627
Attrib TF_COMP      0.40065567514874706
Attrib TF_ZCOMB     -0.4013891992518487
Attrib TF_ZPAIR     0.30552439934550013
Attrib TF_ZSURF     -0.7431254141410357
Attrib TF_PP        0.08311907573957673
Attrib TF_COABS     -0.8539611555012772
Attrib TF_COREL     0.8925272481810163
Attrib TF_RG        0.38230360814687936
Sigmoid Node 4
  Inputs  Weights
  Threshold -0.41263010103926084
  Attrib ALI_SEQIDE  0.13992916966016372
  Attrib ALI_ZSCO   -0.5080159691930051
  Attrib ALI_PBEVAL -0.1895806155000029
  Attrib ALI_TGCOV  -0.1760299103415262
  Attrib ML_LEN     -0.5548275817917732
  Attrib ML_COMP    -0.17749401110628166
  Attrib ML_ZCOMB   -0.2765963229553869
  Attrib ML_ZPAIR   -0.14781053706365996
  Attrib ML_ZSURF   -0.3807260846829993
  Attrib ML_PP      -0.6311156400352186
  Attrib ML_COABS   0.1792586383957213
  Attrib ML_COREL   -0.03956385987122658
  Attrib ML_RG      -0.2805582176013339
  Attrib TF_LEN     0.059804346266281096
  Attrib TF_COMP    -0.11700860531513055
  Attrib TF_ZCOMB   0.265738478494089
  Attrib TF_ZPAIR   0.057966791912817145
  Attrib TF_ZSURF   0.2807089724201649
  Attrib TF_PP      0.006343733601513372
  Attrib TF_COABS   0.5497067650086536
  Attrib TF_COREL   -0.05564195906251746
  Attrib TF_RG      -0.18705002613335442
Sigmoid Node 5
  Inputs  Weights
  Threshold -0.44567454625326514
  Attrib ALI_SEQIDE -0.18819450129907822
  Attrib ALI_ZSCO   -0.38286930908498096
  Attrib ALI_PBEVAL -0.48004787892603706
  Attrib ALI_TGCOV  -0.10587181817006212
  Attrib ML_LEN     -1.1320025725933607
  Attrib ML_COMP    0.5989052437803414
  Attrib ML_ZCOMB   -0.2984219355193872
  Attrib ML_ZPAIR   0.027913490728086413
  Attrib ML_ZSURF   -0.5154060545028943
  Attrib ML_PP      -0.9418832677869162
  Attrib ML_COABS   0.1338514365016201
  Attrib ML_COREL   0.32521298505226737
  Attrib ML_RG      -0.6318729294303737
  Attrib TF_LEN     -0.115740696663662008
  Attrib TF_COMP    0.6193944282775876
  Attrib TF_ZCOMB   0.2931033833376238
  Attrib TF_ZPAIR   0.2351443773610093
  Attrib TF_ZSURF   0.07790945927033542
  Attrib TF_PP      -0.07053145034892884
  Attrib TF_COABS   0.6409569077192196
  Attrib TF_COREL   0.233503799816354
  Attrib TF_RG      -0.41439492862217336
Sigmoid Node 6
  Inputs  Weights
  Threshold -0.18378638984450596
  Attrib ALI_SEQIDE  0.04685564181310289
  Attrib ALI_ZSCO   -0.2053971921524182
  Attrib ALI_PBEVAL -0.02479344673031701
  Attrib ALI_TGCOV  -0.03366509934232431
  Attrib ML_LEN     -0.24397045936883568
  Attrib ML_COMP    -0.03283583975368476
  Attrib ML_ZCOMB   -0.08725047254069915
  Attrib ML_ZPAIR   -0.04049845182412099
  Attrib ML_ZSURF   -0.14913731171443567
  Attrib ML_PP      -0.24409964907753073
  Attrib ML_COABS   0.0822381659610705
  Attrib ML_COREL   0.047131428063602585

```

```

Attrib ML_RG      -0.05628703655480133
Attrib TF_LEN     -0.03372767424967247
Attrib TF_COMP    0.039364321170746126
Attrib TF_ZCOMB   0.06852089847904416
Attrib TF_ZPAIR   0.04079023554577513
Attrib TF_ZSURF   0.08561087682408883
Attrib TF_PP      0.031091653309047732
Attrib TF_COABS   0.24083641712835574
Attrib TF_COREL   0.05383169314149935
Attrib TF_RG      0.016670295165144663
Sigmoid Node 7
  Inputs      Weights
  Threshold   0.4334011595767599
  Attrib ALI_SEQIDE 0.3562457497483074
  Attrib ALI_ZSCO  0.3767165642233656
  Attrib ALI_PBEVAL 0.5591548560216114
  Attrib ALI_TGCOV 0.03611617759138512
  Attrib ML_LEN    1.1338766959956712
  Attrib ML_COMP   -0.6202318951689653
  Attrib ML_ZCOMB  0.24140051024139064
  Attrib ML_ZPAIR  -0.013641827254220222
  Attrib ML_ZSURF  0.4377593256549099
  Attrib ML_PP     0.862387012336021
  Attrib ML_COABS  -0.15216134481674065
  Attrib ML_COREL  -0.3712618548970232
  Attrib ML_RG     0.6928845495654915
  Attrib TF_LEN    0.13212534045448082
  Attrib TF_COMP   -0.6268630537410065
  Attrib TF_ZCOMB  -0.3363385414234979
  Attrib TF_ZPAIR  -0.3279002711695404
  Attrib TF_ZSURF  -0.08512571143505286
  Attrib TF_PP     0.10818297020217536
  Attrib TF_COABS  -0.7581856998044281
  Attrib TF_COREL  -0.2949676736415843
  Attrib TF_RG     0.562319594604412
Sigmoid Node 8
  Inputs      Weights
  Threshold   -1.2552578161051748
  Attrib ALI_SEQIDE 0.18736931649896418
  Attrib ALI_ZSCO  -1.1790510992042167
  Attrib ALI_PBEVAL -0.3695239374498299
  Attrib ALI_TGCOV -0.4676942908664058
  Attrib ML_LEN    -1.5717279832391262
  Attrib ML_COMP   0.00918938586212089
  Attrib ML_ZCOMB  -0.6525899222176728
  Attrib ML_ZPAIR  -0.20216212866131222
  Attrib ML_ZSURF  -0.8316522679316389
  Attrib ML_PP     -1.438866439944451
  Attrib ML_COABS  0.1651160024419076
  Attrib ML_COREL  -0.06808533415999002
  Attrib ML_RG     -0.5099472600488137
  Attrib TF_LEN    0.13499412770628236
  Attrib TF_COMP   0.11417050572634312
  Attrib TF_ZCOMB  0.5038360414063807
  Attrib TF_ZPAIR  -0.028679332137072862
  Attrib TF_ZSURF  0.6116126525398524
  Attrib TF_PP     0.14248730975281784
  Attrib TF_COABS  1.1673670620523855
  Attrib TF_COREL  -0.2286998709715289
  Attrib TF_RG     -0.28654767984567336
Sigmoid Node 9
  Inputs      Weights
  Threshold   -0.8166442575888713
  Attrib ALI_SEQIDE -0.17401739274059844
  Attrib ALI_ZSCO   -0.7783976338276737
  Attrib ALI_PBEVAL -0.548278327632092
  Attrib ALI_TGCOV  -0.22210182003227633
  Attrib ML_LEN     -1.4210633879491466
  Attrib ML_COMP    0.40264919895768947
  Attrib ML_ZCOMB   -0.5043492865010478
  Attrib ML_ZPAIR   -0.10274028689011669
  Attrib ML_ZSURF   -0.6607839530589311
  Attrib ML_PP      -1.2541209223865057
  Attrib ML_COABS   0.19939595818467945
  Attrib ML_COREL   0.15994577580095468
  Attrib ML_RG      -0.641497228895988
  Attrib TF_LEN     -0.02873342613822372

```


Attrib TF_COMP	0.5321181275181656
Attrib TF_ZCOMB	0.3797212387299009
Attrib TF_ZPAIR	0.1699504442698747
Attrib TF_ZSURF	0.3561032350609874
Attrib TF_PP	0.03978635727029743
Attrib TF_COABS	0.9926159863113616
Attrib TF_COREL	-0.007522620906124857
Attrib TF_RG	-0.4999343965375808

Sigmoid Node 10

Inputs	Weights
Threshold	1.5040010570451734
Attrib ALI_SEQIDE	-0.3278228980612083
Attrib ALI_ZSCO	1.3543755611846988
Attrib ALI_PBEVAL	0.29708080114775565
Attrib ALI_TGCOV	0.8187630281073842
Attrib ML_LEN	1.8125729162973983
Attrib ML_COMP	0.33845465704073396
Attrib ML_ZCOMB	0.6982081161091078
Attrib ML_ZPAIR	0.2547134619769126
Attrib ML_ZSURF	0.8317982762182989
Attrib ML_PP	1.5823574313945385
Attrib ML_COABS	-0.21840566362119476
Attrib ML_COREL	0.12785826832710975
Attrib ML_RG	0.4527633990288584
Attrib TF_LEN	-0.09593595477108886
Attrib TF_COMP	0.09134928394425436
Attrib TF_ZCOMB	-0.6559673015358221
Attrib TF_ZPAIR	0.029379330212791054
Attrib TF_ZSURF	-0.8119492826808699
Attrib TF_PP	-0.2786054569274651
Attrib TF_COABS	-1.3289028336693414
Attrib TF_COREL	0.3096425922982166
Attrib TF_RG	0.19005841104983107

Sigmoid Node 11

Inputs	Weights
Threshold	0.2718326747599342
Attrib ALI_SEQIDE	0.1665056049129975
Attrib ALI_ZSCO	0.3281432341692342
Attrib ALI_PBEVAL	0.4903719507163219
Attrib ALI_TGCOV	0.10493787884627061
Attrib ML_LEN	0.8195793655069079
Attrib ML_COMP	-0.07590217191010824
Attrib ML_ZCOMB	0.21538284805436053
Attrib ML_ZPAIR	0.021609185700196657
Attrib ML_ZSURF	0.3975331415305279
Attrib ML_PP	0.7594282155035249
Attrib ML_COABS	-0.15771776414631009
Attrib ML_COREL	-0.15137658405626989
Attrib ML_RG	0.4555749953545961
Attrib TF_LEN	0.037478804016601745
Attrib TF_COMP	-0.2103545982461806
Attrib TF_ZCOMB	-0.2895508542685876
Attrib TF_ZPAIR	-0.1999180384429176
Attrib TF_ZSURF	-0.18119520098628014
Attrib TF_PP	0.026061409869113622
Attrib TF_COABS	-0.5937469498208277
Attrib TF_COREL	-0.07342970355405228
Attrib TF_RG	0.3523217597913659

Sigmoid Node 12

Inputs	Weights
Threshold	-1.003847526873705
Attrib ALI_SEQIDE	-0.39466458226738027
Attrib ALI_ZSCO	-0.8169463874583853
Attrib ALI_PBEVAL	-0.881295796956289
Attrib ALI_TGCOV	-0.20421683612878472
Attrib ML_LEN	-1.7955964983723451
Attrib ML_COMP	1.0832613064315229
Attrib ML_ZCOMB	-0.5719196425665912
Attrib ML_ZPAIR	-0.04772644022208062
Attrib ML_ZSURF	-0.9201647923182323
Attrib ML_PP	-1.470149020666726
Attrib ML_COABS	0.22194525970256204
Attrib ML_COREL	0.42212832352827534
Attrib ML_RG	-0.8550413410739625
Attrib TF_LEN	-0.08814027770875076
Attrib TF_COMP	1.1769096415283118
Attrib TF_ZCOMB	0.47480157906194626

```

Attrib TF_ZPAIR      0.306338468400136
Attrib TF_ZSURF     0.24843139723153
Attrib TF_PP        -0.023192461790702106
Attrib TF_COABS     1.1559935343382615
Attrib TF_COREL     0.17560928522014352
Attrib TF_RG        -0.6850459554697669
Sigmoid Node 13
  Inputs      Weights
Threshold    1.7140828834805266
Attrib ALI_SEQIDE  0.09814727711107703
Attrib ALI_ZSCO   2.1811269172945305
Attrib ALI_PBEVAL 0.7391871543507997
Attrib ALI_TGCOV  -1.20938445691984
Attrib ML_LEN     1.4785520985648182
Attrib ML_COMP    -0.13288337571930445
Attrib ML_ZCOMB   1.4272118600667476
Attrib ML_ZPAIR   0.5287580697230914
Attrib ML_ZSURF   1.818874337401895
Attrib ML_PP      1.591228373936003
Attrib ML_COABS   0.017853502978932914
Attrib ML_COREL   0.48301601196472665
Attrib ML_RG      0.19875587940410525
Attrib TF_LEN     -0.5703586982769134
Attrib TF_COMP    -0.6197275000323512
Attrib TF_ZCOMB   0.2382041025775588
Attrib TF_ZPAIR   0.6628280627781283
Attrib TF_ZSURF   0.13821739377497907
Attrib TF_PP      -0.5177510276510243
Attrib TF_COABS   -1.129326785699243
Attrib TF_COREL   0.7201438329537515
Attrib TF_RG      0.0550146217989861
Class 1
  Input
  Node 0
Class 0
  Input
  Node 1

```

Anexo H: intervalos de confianza de la diferencia de las AUC para cada par de variables como clasificadores.

TEST1/TEST2	AUC_DIFFERENCE	CONFIDENCE_INTERVAL
"ALI_SEQIDE"/"ALI_ZSCO"	-0.0791752	(-0.106147 , -0.0522035)
"ALI_SEQIDE"/"ALI_PBEVAL"	0.0289322	(-0.00760116 , 0.0654656)
"ALI_SEQIDE"/"ALI_TGCOV"	0.207299	(0.163178 , 0.251421)
"ALI_SEQIDE"/"ML_LEN"	0.253545	(0.208841 , 0.29825)
"ALI_SEQIDE"/"ML_COMP"	0.083405	(0.0442322 , 0.122578)
"ALI_SEQIDE"/"ML_ZCOMB"	-0.0821554	(-0.115863 , -0.0484475)
"ALI_SEQIDE"/"ML_ZPAIR"	-0.0447273	(-0.0795852 , -0.00986931)
"ALI_SEQIDE"/"ML_ZSURF"	-0.0769802	(-0.111307 , -0.0426532)
"ALI_SEQIDE"/"ML_PP"	-0.0612512	(-0.0953651 , -0.0271374)
"ALI_SEQIDE"/"ML_COABS"	0.0290116	(-0.0124807 , 0.0705039)
"ALI_SEQIDE"/"ML_COREL"	-0.0207355	(-0.0592386 , 0.0177676)
"ALI_SEQIDE"/"ML_RG"	0.110412	(0.0691306 , 0.151693)
"ALI_SEQIDE"/"TF_LEN"	0.232617	(0.18764 , 0.277593)
"ALI_SEQIDE"/"TF_COMP"	0.0921339	(0.0530012 , 0.131267)
"ALI_SEQIDE"/"TF_ZCOMB"	0.0992463	(0.0564414 , 0.142051)
"ALI_SEQIDE"/"TF_ZPAIR"	0.1686	(0.124753 , 0.212447)
"ALI_SEQIDE"/"TF_ZSURF"	0.043714	(0.00273734 , 0.0846908)
"ALI_SEQIDE"/"TF_PP"	0.191202	(0.149747 , 0.232656)
"ALI_SEQIDE"/"TF_COABS"	0.00539174	(-0.0352041 , 0.0459876)
"ALI_SEQIDE"/"TF_COREL"	-0.0221587	(-0.060307 , 0.0159897)
"ALI_SEQIDE"/"TF_RG"	0.11985	(0.0786312 , 0.161068)
"ALI_SEQIDE"/"TW_LEN"	0.24738	(0.202087 , 0.292673)
"ALI_SEQIDE"/"TW_COMP"	0.102572	(0.0629854 , 0.142158)
"ALI_SEQIDE"/"TW_ZCOMB"	0.121033	(0.0778298 , 0.164236)
"ALI_SEQIDE"/"TW_ZPAIR"	0.191812	(0.147788 , 0.235835)
"ALI_SEQIDE"/"TW_ZSURF"	0.0577851	(0.0160717 , 0.0994986)
"ALI_SEQIDE"/"TW_PP"	0.21495	(0.173795 , 0.256106)
"ALI_SEQIDE"/"TW_COABS"	0.0194364	(-0.0220932 , 0.060966)
"ALI_SEQIDE"/"TW_COREL"	-0.0228463	(-0.0608205 , 0.0151279)
"ALI_SEQIDE"/"TW_RG"	0.134686	(0.0934495 , 0.175922)
"ALI_ZSCO"/"ALI_PBEVAL"	0.108107	(0.0817543 , 0.134461)
"ALI_ZSCO"/"ALI_TGCOV"	0.286474	(0.251205 , 0.321744)
"ALI_ZSCO"/"ML_LEN"	0.332721	(0.302364 , 0.363077)
"ALI_ZSCO"/"ML_COMP"	0.16258	(0.122462 , 0.202699)
"ALI_ZSCO"/"ML_ZCOMB"	-0.00298017	(-0.0251898 , 0.0192294)
"ALI_ZSCO"/"ML_ZPAIR"	0.0344479	(0.00738652 , 0.0615093)
"ALI_ZSCO"/"ML_ZSURF"	0.00219504	(-0.018974 , 0.023364)
"ALI_ZSCO"/"ML_PP"	0.017924	(-0.0148309 , 0.0506788)
"ALI_ZSCO"/"ML_COABS"	0.108187	(0.0813634 , 0.13501)
"ALI_ZSCO"/"ML_COREL"	0.0584397	(0.0227993 , 0.09408)
"ALI_ZSCO"/"ML_RG"	0.189587	(0.142974 , 0.236199)
"ALI_ZSCO"/"TF_LEN"	0.311792	(0.281479 , 0.342105)
"ALI_ZSCO"/"TF_COMP"	0.171309	(0.13081 , 0.211808)
"ALI_ZSCO"/"TF_ZCOMB"	0.178421	(0.147815 , 0.209028)
"ALI_ZSCO"/"TF_ZPAIR"	0.247775	(0.212904 , 0.282646)
"ALI_ZSCO"/"TF_ZSURF"	0.122889	(0.096037 , 0.149742)
"ALI_ZSCO"/"TF_PP"	0.270377	(0.228375 , 0.312379)
"ALI_ZSCO"/"TF_COABS"	0.0845669	(0.0585956 , 0.110538)
"ALI_ZSCO"/"TF_COREL"	0.0570165	(0.0217449 , 0.0922882)
"ALI_ZSCO"/"TF_RG"	0.199025	(0.152498 , 0.245552)
"ALI_ZSCO"/"TW_LEN"	0.326555	(0.296037 , 0.357074)
"ALI_ZSCO"/"TW_COMP"	0.181747	(0.141051 , 0.222443)
"ALI_ZSCO"/"TW_ZCOMB"	0.200208	(0.169275 , 0.231142)
"ALI_ZSCO"/"TW_ZPAIR"	0.270987	(0.236608 , 0.305365)
"ALI_ZSCO"/"TW_ZSURF"	0.13696	(0.109422 , 0.164499)
"ALI_ZSCO"/"TW_PP"	0.294126	(0.251845 , 0.336406)
"ALI_ZSCO"/"TW_COABS"	0.0986116	(0.0721167 , 0.125106)
"ALI_ZSCO"/"TW_COREL"	0.0563289	(0.0210778 , 0.09158)
"ALI_ZSCO"/"TW_RG"	0.213861	(0.167026 , 0.260696)
"ALI_PBEVAL"/"ALI_TGCOV"	0.178367	(0.147237 , 0.209497)
"ALI_PBEVAL"/"ML_LEN"	0.224613	(0.194569 , 0.254657)
"ALI_PBEVAL"/"ML_COMP"	0.0544727	(0.0146049 , 0.0943405)
"ALI_PBEVAL"/"ML_ZCOMB"	-0.111088	(-0.136093 , -0.086082)
"ALI_PBEVAL"/"ML_ZPAIR"	-0.0736595	(-0.100024 , -0.047295)
"ALI_PBEVAL"/"ML_ZSURF"	-0.105912	(-0.133104 , -0.0787213)
"ALI_PBEVAL"/"ML_PP"	-0.0901835	(-0.127727 , -0.0526399)
"ALI_PBEVAL"/"ML_COABS"	7.93388e-05	(-0.0317009 , 0.0318596)
"ALI_PBEVAL"/"ML_COREL"	-0.0496678	(-0.0923432 , -0.00699236)
"ALI_PBEVAL"/"ML_RG"	0.0814793	(0.032031 , 0.130928)
"ALI_PBEVAL"/"TF_LEN"	0.203684	(0.174354 , 0.233015)

"ALI_PBEVAL" / "TF_COMP"	0.0632017	(0.0238013 , 0.102602)
"ALI_PBEVAL" / "TF_ZCOMB"	0.070314	(0.0409938 , 0.0996343)
"ALI_PBEVAL" / "TF_ZPAIR"	0.139668	(0.105935 , 0.173401)
"ALI_PBEVAL" / "TF_ZSURF"	0.0147818	(-0.0129721 , 0.0425358)
"ALI_PBEVAL" / "TF_PP"	0.162269	(0.117365 , 0.207173)
"ALI_PBEVAL" / "TF_COABS"	-0.0235405	(-0.0542155 , 0.00713448)
"ALI_PBEVAL" / "TF_COREL"	-0.0510909	(-0.0935354 , -0.00864639)
"ALI_PBEVAL" / "TF_RG"	0.0909174	(0.0421578 , 0.139677)
"ALI_PBEVAL" / "TW_LEN"	0.218448	(0.188384 , 0.248512)
"ALI_PBEVAL" / "TW_COMP"	0.0736397	(0.0336458 , 0.113634)
"ALI_PBEVAL" / "TW_ZCOMB"	0.0921008	(0.0615462 , 0.122655)
"ALI_PBEVAL" / "TW_ZPAIR"	0.162879	(0.129129 , 0.19663)
"ALI_PBEVAL" / "TW_ZSURF"	0.0288529	(-0.000160982 , 0.0578668)
"ALI_PBEVAL" / "TW_PP"	0.186018	(0.141203 , 0.230833)
"ALI_PBEVAL" / "TW_COABS"	-0.00949587	(-0.0408858 , 0.0218941)
"ALI_PBEVAL" / "TW_COREL"	-0.0517785	(-0.0941538 , -0.00940327)
"ALI_PBEVAL" / "TW_RG"	0.105754	(0.0568386 , 0.154669)
"ALI_TGCOV" / "ML_LEN"	0.0462463	(0.0120416 , 0.0804509)
"ALI_TGCOV" / "ML_COMP"	-0.123894	(-0.166459 , -0.0813292)
"ALI_TGCOV" / "ML_ZCOMB"	-0.289455	(-0.323325 , -0.255584)
"ALI_TGCOV" / "ML_ZPAIR"	-0.252026	(-0.287642 , -0.216411)
"ALI_TGCOV" / "ML_ZSURF"	-0.284279	(-0.318449 , -0.25011)
"ALI_TGCOV" / "ML_PP"	-0.26855	(-0.309404 , -0.227697)
"ALI_TGCOV" / "ML_COABS"	-0.178288	(-0.214398 , -0.142177)
"ALI_TGCOV" / "ML_COREL"	-0.228035	(-0.273704 , -0.182365)
"ALI_TGCOV" / "ML_RG"	-0.0968876	(-0.149953 , -0.0438223)
"ALI_TGCOV" / "TF_LEN"	0.0253174	(-0.00872642 , 0.0593611)
"ALI_TGCOV" / "TF_COMP"	-0.115165	(-0.158264 , -0.0720661)
"ALI_TGCOV" / "TF_ZCOMB"	-0.108053	(-0.14257 , -0.0735363)
"ALI_TGCOV" / "TF_ZPAIR"	-0.0386992	(-0.0764375 , -0.000960896)
"ALI_TGCOV" / "TF_ZSURF"	-0.163585	(-0.197494 , -0.129676)
"ALI_TGCOV" / "TF_PP"	-0.0160975	(-0.0655378 , 0.0333427)
"ALI_TGCOV" / "TF_COABS"	-0.201907	(-0.237431 , -0.166384)
"ALI_TGCOV" / "TF_COREL"	-0.229458	(-0.274813 , -0.184103)
"ALI_TGCOV" / "TF_RG"	-0.0874496	(-0.139222 , -0.0356772)
"ALI_TGCOV" / "TW_LEN"	0.040081	(0.00602339 , 0.0741386)
"ALI_TGCOV" / "TW_COMP"	-0.104727	(-0.147928 , -0.0615261)
"ALI_TGCOV" / "TW_ZCOMB"	-0.0862661	(-0.121309 , -0.0512229)
"ALI_TGCOV" / "TW_ZPAIR"	-0.0154876	(-0.0534277 , 0.0224525)
"ALI_TGCOV" / "TW_ZSURF"	-0.149514	(-0.183709 , -0.115319)
"ALI_TGCOV" / "TW_PP"	0.00765124	(-0.0417697 , 0.0570722)
"ALI_TGCOV" / "TW_COABS"	-0.187863	(-0.224213 , -0.151512)
"ALI_TGCOV" / "TW_COREL"	-0.230145	(-0.275745 , -0.184546)
"ALI_TGCOV" / "TW_RG"	-0.0726132	(-0.124522 , -0.0207044)
"ML_LEN" / "ML_COMP"	-0.17014	(-0.216194 , -0.124087)
"ML_LEN" / "ML_ZCOMB"	-0.335701	(-0.36552 , -0.305881)
"ML_LEN" / "ML_ZPAIR"	-0.298273	(-0.329914 , -0.266631)
"ML_LEN" / "ML_ZSURF"	-0.330526	(-0.361115 , -0.299937)
"ML_LEN" / "ML_PP"	-0.314797	(-0.356177 , -0.273417)
"ML_LEN" / "ML_COABS"	-0.224534	(-0.249706 , -0.199362)
"ML_LEN" / "ML_COREL"	-0.274281	(-0.323772 , -0.22479)
"ML_LEN" / "ML_RG"	-0.143134	(-0.204528 , -0.0817397)
"ML_LEN" / "TF_LEN"	-0.0209289	(-0.0267965 , -0.0150614)
"ML_LEN" / "TF_COMP"	-0.161412	(-0.207312 , -0.115511)
"ML_LEN" / "TF_ZCOMB"	-0.154299	(-0.180936 , -0.127663)
"ML_LEN" / "TF_ZPAIR"	-0.0849455	(-0.115096 , -0.0547949)
"ML_LEN" / "TF_ZSURF"	-0.209831	(-0.237637 , -0.182025)
"ML_LEN" / "TF_PP"	-0.0623438	(-0.110545 , -0.0141431)
"ML_LEN" / "TF_COABS"	-0.248154	(-0.273749 , -0.222558)
"ML_LEN" / "TF_COREL"	-0.275704	(-0.324794 , -0.226614)
"ML_LEN" / "TF_RG"	-0.133696	(-0.193851 , -0.0735405)
"ML_LEN" / "TW_LEN"	-0.00616529	(-0.0154484 , 0.00311781)
"ML_LEN" / "TW_COMP"	-0.150974	(-0.197289 , -0.104658)
"ML_LEN" / "TW_ZCOMB"	-0.132512	(-0.159261 , -0.105764)
"ML_LEN" / "TW_ZPAIR"	-0.0617339	(-0.091583 , -0.0318847)
"ML_LEN" / "TW_ZSURF"	-0.19576	(-0.223289 , -0.168232)
"ML_LEN" / "TW_PP"	-0.038595	(-0.0865809 , 0.00939078)
"ML_LEN" / "TW_COABS"	-0.234109	(-0.260024 , -0.208194)
"ML_LEN" / "TW_COREL"	-0.276392	(-0.325611 , -0.227172)
"ML_LEN" / "TW_RG"	-0.11886	(-0.179051 , -0.0586678)
"ML_COMP" / "ML_ZCOMB"	-0.16556	(-0.201329 , -0.129792)
"ML_COMP" / "ML_ZPAIR"	-0.128132	(-0.164397 , -0.0918674)
"ML_COMP" / "ML_ZSURF"	-0.160385	(-0.19771 , -0.123061)
"ML_COMP" / "ML_PP"	-0.144656	(-0.18133 , -0.107982)
"ML_COMP" / "ML_COABS"	-0.0543934	(-0.0964602 , -0.0123266)
"ML_COMP" / "ML_COREL"	-0.10414	(-0.142009 , -0.0662717)
"ML_COMP" / "ML_RG"	0.0270066	(-0.00305984 , 0.0570731)

"ML_COMP" / "TF_LEN"	0.149212	(0.103475 , 0.194948)
"ML_COMP" / "TF_COMP"	0.00872893	(-0.00726986 , 0.0247277)
"ML_COMP" / "TF_ZCOMB"	0.0158413	(-0.0248768 , 0.0565595)
"ML_COMP" / "TF_ZPAIR"	0.085195	(0.0435545 , 0.126836)
"ML_COMP" / "TF_ZSURF"	-0.0396909	(-0.0793966 , 1.48205e-05)
"ML_COMP" / "TF_PP"	0.107797	(0.0663616 , 0.149232)
"ML_COMP" / "TF_COABS"	-0.0780132	(-0.119108 , -0.0369188)
"ML_COMP" / "TF_COREL"	-0.105564	(-0.14372 , -0.0674073)
"ML_COMP" / "TF_RG"	0.0364446	(0.00624072 , 0.0666485)
"ML_COMP" / "TW_LEN"	0.163975	(0.117819 , 0.210131)
"ML_COMP" / "TW_COMP"	0.0191669	(0.00277145 , 0.0355624)
"ML_COMP" / "TW_ZCOMB"	0.0376281	(-0.00378513 , 0.0790413)
"ML_COMP" / "TW_ZPAIR"	0.108407	(0.0662494 , 0.150564)
"ML_COMP" / "TW_ZSURF"	-0.0256198	(-0.0657402 , 0.0145005)
"ML_COMP" / "TW_PP"	0.131545	(0.0904625 , 0.172628)
"ML_COMP" / "TW_COABS"	-0.0639686	(-0.105956 , -0.0219815)
"ML_COMP" / "TW_COREL"	-0.106251	(-0.144345 , -0.0681578)
"ML_COMP" / "TW_RG"	0.051281	(0.0212023 , 0.0813596)
"ML_ZCOMB" / "ML_ZPAIR"	0.0374281	(0.0247278 , 0.0501283)
"ML_ZCOMB" / "ML_ZSURF"	0.00517521	(-0.00634602 , 0.0166964)
"ML_ZCOMB" / "ML_PP"	0.0209041	(-0.00836691 , 0.0501752)
"ML_ZCOMB" / "ML_COABS"	0.111167	(0.0833941 , 0.13894)
"ML_ZCOMB" / "ML_COREL"	0.0614198	(0.0256159 , 0.0972237)
"ML_ZCOMB" / "ML_RG"	0.192567	(0.147554 , 0.23758)
"ML_ZCOMB" / "TF_LEN"	0.314772	(0.285449 , 0.344095)
"ML_ZCOMB" / "TF_COMP"	0.174289	(0.138484 , 0.210094)
"ML_ZCOMB" / "TF_ZCOMB"	0.181402	(0.155325 , 0.207479)
"ML_ZCOMB" / "TF_ZPAIR"	0.250755	(0.22001 , 0.281501)
"ML_ZCOMB" / "TF_ZSURF"	0.125869	(0.102798 , 0.14894)
"ML_ZCOMB" / "TF_PP"	0.273357	(0.234813 , 0.311902)
"ML_ZCOMB" / "TF_COABS"	0.0875471	(0.0608299 , 0.114264)
"ML_ZCOMB" / "TF_COREL"	0.0599967	(0.0242976 , 0.0956958)
"ML_ZCOMB" / "TF_RG"	0.202005	(0.157712 , 0.246298)
"ML_ZCOMB" / "TW_LEN"	0.329536	(0.299788 , 0.359283)
"ML_ZCOMB" / "TW_COMP"	0.184727	(0.148621 , 0.220833)
"ML_ZCOMB" / "TW_ZCOMB"	0.203188	(0.17608 , 0.230297)
"ML_ZCOMB" / "TW_ZPAIR"	0.273967	(0.243133 , 0.304801)
"ML_ZCOMB" / "TW_ZSURF"	0.13994	(0.11585 , 0.164031)
"ML_ZCOMB" / "TW_PP"	0.297106	(0.258251 , 0.335961)
"ML_ZCOMB" / "TW_COABS"	0.101592	(0.0739309 , 0.129253)
"ML_ZCOMB" / "TW_COREL"	0.0593091	(0.0235025 , 0.0951157)
"ML_ZCOMB" / "TW_RG"	0.216841	(0.172296 , 0.261386)
"ML_ZPAIR" / "ML_ZSURF"	-0.0322529	(-0.0546436 , -0.00986223)
"ML_ZPAIR" / "ML_PP"	-0.016524	(-0.0468818 , 0.0138339)
"ML_ZPAIR" / "ML_COABS"	0.0737388	(0.0425895 , 0.104888)
"ML_ZPAIR" / "ML_COREL"	0.0239917	(-0.0140847 , 0.0620681)
"ML_ZPAIR" / "ML_RG"	0.155139	(0.109344 , 0.200934)
"ML_ZPAIR" / "TF_LEN"	0.277344	(0.24612 , 0.308568)
"ML_ZPAIR" / "TF_COMP"	0.136861	(0.101157 , 0.172566)
"ML_ZPAIR" / "TF_ZCOMB"	0.143974	(0.114985 , 0.172962)
"ML_ZPAIR" / "TF_ZPAIR"	0.213327	(0.181389 , 0.245265)
"ML_ZPAIR" / "TF_ZSURF"	0.0884413	(0.0606262 , 0.116256)
"ML_ZPAIR" / "TF_PP"	0.235929	(0.197226 , 0.274631)
"ML_ZPAIR" / "TF_COABS"	0.050119	(0.0199714 , 0.0802666)
"ML_ZPAIR" / "TF_COREL"	0.0225686	(-0.0153817 , 0.0605189)
"ML_ZPAIR" / "TF_RG"	0.164577	(0.119822 , 0.209332)
"ML_ZPAIR" / "TW_LEN"	0.292107	(0.260472 , 0.323742)
"ML_ZPAIR" / "TW_COMP"	0.147299	(0.111201 , 0.183398)
"ML_ZPAIR" / "TW_ZCOMB"	0.16576	(0.13601 , 0.195511)
"ML_ZPAIR" / "TW_ZPAIR"	0.236539	(0.20447 , 0.268608)
"ML_ZPAIR" / "TW_ZSURF"	0.102512	(0.0739289 , 0.131096)
"ML_ZPAIR" / "TW_PP"	0.259678	(0.220892 , 0.298463)
"ML_ZPAIR" / "TW_COABS"	0.0641636	(0.0331177 , 0.0952095)
"ML_ZPAIR" / "TW_COREL"	0.021881	(-0.0162519 , 0.0600139)
"ML_ZPAIR" / "TW_RG"	0.179413	(0.134567 , 0.224259)
"ML_ZSURF" / "ML_PP"	0.0157289	(-0.0146346 , 0.0460924)
"ML_ZSURF" / "ML_COABS"	0.105992	(0.0782646 , 0.133719)
"ML_ZSURF" / "ML_COREL"	0.0562446	(0.0203766 , 0.0921126)
"ML_ZSURF" / "ML_RG"	0.187392	(0.141716 , 0.233068)
"ML_ZSURF" / "TF_LEN"	0.309597	(0.279414 , 0.33978)
"ML_ZSURF" / "TF_COMP"	0.169114	(0.131254 , 0.206974)
"ML_ZSURF" / "TF_ZCOMB"	0.176226	(0.149447 , 0.203006)
"ML_ZSURF" / "TF_ZPAIR"	0.24558	(0.213544 , 0.277616)
"ML_ZSURF" / "TF_ZSURF"	0.120694	(0.0979885 , 0.1434)
"ML_ZSURF" / "TF_PP"	0.268182	(0.227857 , 0.308507)
"ML_ZSURF" / "TF_COABS"	0.0823719	(0.0555307 , 0.109213)
"ML_ZSURF" / "TF_COREL"	0.0548215	(0.019027 , 0.0906159)

"ML_ZSURF" / "TF_RG"	0.19683	(0.15157 , 0.242089)
"ML_ZSURF" / "TW_LEN"	0.32436	(0.293732 , 0.354988)
"ML_ZSURF" / "TW_COMP"	0.179552	(0.141465 , 0.217639)
"ML_ZSURF" / "TW_ZCOMB"	0.198013	(0.17018 , 0.225847)
"ML_ZSURF" / "TW_ZPAIR"	0.268792	(0.236632 , 0.300951)
"ML_ZSURF" / "TW_ZSURF"	0.134765	(0.11096 , 0.15857)
"ML_ZSURF" / "TW_PP"	0.291931	(0.251214 , 0.332647)
"ML_ZSURF" / "TW_COABS"	0.0964165	(0.0686569 , 0.124176)
"ML_ZSURF" / "TW_COREL"	0.0541339	(0.0183299 , 0.0899379)
"ML_ZSURF" / "TW_RG"	0.211666	(0.16608 , 0.257253)
"ML_PP" / "ML_COABS"	0.0902628	(0.0523845 , 0.128141)
"ML_PP" / "ML_COREL"	0.0405157	(0.00529188 , 0.0757395)
"ML_PP" / "ML_RG"	0.171663	(0.131681 , 0.211645)
"ML_PP" / "TF_LEN"	0.293868	(0.252856 , 0.33488)
"ML_PP" / "TF_COMP"	0.153385	(0.116269 , 0.190502)
"ML_PP" / "TF_ZCOMB"	0.160498	(0.121986 , 0.199009)
"ML_PP" / "TF_ZPAIR"	0.229851	(0.190304 , 0.269398)
"ML_PP" / "TF_ZSURF"	0.104965	(0.0681467 , 0.141784)
"ML_PP" / "TF_PP"	0.252453	(0.218607 , 0.286299)
"ML_PP" / "TF_COABS"	0.066643	(0.0299679 , 0.103318)
"ML_PP" / "TF_COREL"	0.0390926	(0.00386949 , 0.0743156)
"ML_PP" / "TF_RG"	0.181101	(0.141505 , 0.220697)
"ML_PP" / "TW_LEN"	0.308631	(0.267562 , 0.349701)
"ML_PP" / "TW_COMP"	0.163823	(0.126513 , 0.201134)
"ML_PP" / "TW_ZCOMB"	0.182284	(0.143341 , 0.221228)
"ML_PP" / "TW_ZPAIR"	0.253063	(0.21301 , 0.293115)
"ML_PP" / "TW_ZSURF"	0.119036	(0.0822213 , 0.155851)
"ML_PP" / "TW_PP"	0.276202	(0.241736 , 0.310667)
"ML_PP" / "TW_COABS"	0.0806876	(0.0431855 , 0.11819)
"ML_PP" / "TW_COREL"	0.038405	(0.0031599 , 0.07365)
"ML_PP" / "TW_RG"	0.195937	(0.15603 , 0.235844)
"ML_COABS" / "ML_COREL"	-0.0497471	(-0.082896 , -0.0165982)
"ML_COABS" / "ML_RG"	0.0814	(0.0292062 , 0.133594)
"ML_COABS" / "TF_LEN"	0.203605	(0.177921 , 0.229289)
"ML_COABS" / "TF_COMP"	0.0631223	(0.0208839 , 0.105361)
"ML_COABS" / "TF_ZCOMB"	0.0702347	(0.0410837 , 0.0993857)
"ML_COABS" / "TF_ZPAIR"	0.139588	(0.106101 , 0.173076)
"ML_COABS" / "TF_ZSURF"	0.0147025	(-0.0111988 , 0.0406038)
"ML_COABS" / "TF_PP"	0.16219	(0.117524 , 0.206856)
"ML_COABS" / "TF_COABS"	-0.0236198	(-0.029475 , -0.0177647)
"ML_COABS" / "TF_COREL"	-0.0511702	(-0.0840957 , -0.0182447)
"ML_COABS" / "TF_RG"	0.090838	(0.0397818 , 0.141894)
"ML_COABS" / "TW_LEN"	0.218369	(0.192143 , 0.244594)
"ML_COABS" / "TW_COMP"	0.0735603	(0.0309887 , 0.116132)
"ML_COABS" / "TW_ZCOMB"	0.0920215	(0.0628901 , 0.121153)
"ML_COABS" / "TW_ZPAIR"	0.1628	(0.12976 , 0.19584)
"ML_COABS" / "TW_ZSURF"	0.0287736	(0.00275687 , 0.0547902)
"ML_COABS" / "TW_PP"	0.185939	(0.141221 , 0.230657)
"ML_COABS" / "TW_COABS"	-0.00957521	(-0.0177633 , -0.00138709)
"ML_COABS" / "TW_COREL"	-0.0518579	(-0.0851785 , -0.0185372)
"ML_COABS" / "TW_RG"	0.105674	(0.0545336 , 0.156815)
"ML_COREL" / "ML_RG"	0.131147	(0.0977121 , 0.164582)
"ML_COREL" / "TF_LEN"	0.253352	(0.203879 , 0.302825)
"ML_COREL" / "TF_COMP"	0.112869	(0.0738338 , 0.151905)
"ML_COREL" / "TF_ZCOMB"	0.119982	(0.0755279 , 0.164436)
"ML_COREL" / "TF_ZPAIR"	0.189336	(0.142836 , 0.235835)
"ML_COREL" / "TF_ZSURF"	0.0644496	(0.0238662 , 0.105033)
"ML_COREL" / "TF_PP"	0.211937	(0.169942 , 0.253932)
"ML_COREL" / "TF_COABS"	0.0261273	(-0.00607402 , 0.0583286)
"ML_COREL" / "TF_COREL"	-0.00142314	(-0.00738606 , 0.00453978)
"ML_COREL" / "TF_RG"	0.140585	(0.106737 , 0.174433)
"ML_COREL" / "TW_LEN"	0.268116	(0.218727 , 0.317505)
"ML_COREL" / "TW_COMP"	0.123307	(0.0840098 , 0.162605)
"ML_COREL" / "TW_ZCOMB"	0.141769	(0.0967902 , 0.186747)
"ML_COREL" / "TW_ZPAIR"	0.212547	(0.16581 , 0.259284)
"ML_COREL" / "TW_ZSURF"	0.0785207	(0.0373554 , 0.119686)
"ML_COREL" / "TW_PP"	0.235686	(0.193266 , 0.278106)
"ML_COREL" / "TW_COABS"	0.0401719	(0.00699367 , 0.0733501)
"ML_COREL" / "TW_COREL"	-0.00211074	(-0.00900106 , 0.00477957)
"ML_COREL" / "TW_RG"	0.155421	(0.121368 , 0.189475)
"ML_RG" / "TF_LEN"	0.122205	(0.0611637 , 0.183246)
"ML_RG" / "TF_COMP"	-0.0182777	(-0.0506971 , 0.0141417)
"ML_RG" / "TF_ZCOMB"	-0.0111653	(-0.0641219 , 0.0417913)
"ML_RG" / "TF_ZPAIR"	0.0581884	(0.00508944 , 0.111287)
"ML_RG" / "TF_ZSURF"	-0.0666975	(-0.117438 , -0.0159572)
"ML_RG" / "TF_PP"	0.0807901	(0.0375038 , 0.124076)
"ML_RG" / "TF_COABS"	-0.10502	(-0.155884 , -0.054156)

"ML_RG" / "TF_COREL"	-0.13257	(-0.166487 , -0.0986532)
"ML_RG" / "TF_RG"	0.00943802	(-0.00425532 , 0.0231314)
"ML_RG" / "TW_LEN"	0.136969	(0.0759297 , 0.198007)
"ML_RG" / "TW_COMP"	-0.00783967	(-0.0403911 , 0.0247118)
"ML_RG" / "TW_ZCOMB"	0.0106215	(-0.0432382 , 0.0644812)
"ML_RG" / "TW_ZPAIR"	0.0814	(0.0276024 , 0.135198)
"ML_RG" / "TW_ZSURF"	-0.0526264	(-0.104214 , -0.00103853)
"ML_RG" / "TW_PP"	0.104539	(0.0608544 , 0.148223)
"ML_RG" / "TW_COABS"	-0.0909752	(-0.142724 , -0.0392264)
"ML_RG" / "TF_COREL"	-0.133258	(-0.166848 , -0.0996675)
"ML_RG" / "TF_RG"	0.0242744	(0.00920616 , 0.0393426)
"TF_LEN" / "TF_COMP"	-0.140483	(-0.186503 , -0.0944625)
"TF_LEN" / "TF_ZCOMB"	-0.13337	(-0.159539 , -0.107201)
"TF_LEN" / "TF_ZPAIR"	-0.0640165	(-0.0940069 , -0.0340262)
"TF_LEN" / "TF_ZSURF"	-0.188902	(-0.21619 , -0.161615)
"TF_LEN" / "TF_PP"	-0.0414149	(-0.08957 , 0.00674029)
"TF_LEN" / "TF_COABS"	-0.227225	(-0.252676 , -0.201774)
"TF_LEN" / "TF_COREL"	-0.254775	(-0.303926 , -0.205624)
"TF_LEN" / "TF_RG"	-0.112767	(-0.173117 , -0.0524169)
"TF_LEN" / "TW_LEN"	0.0147636	(0.00844794 , 0.0210793)
"TF_LEN" / "TW_COMP"	-0.130045	(-0.176445 , -0.0836441)
"TF_LEN" / "TW_ZCOMB"	-0.111583	(-0.137588 , -0.0855793)
"TF_LEN" / "TW_ZPAIR"	-0.040805	(-0.0701779 , -0.011432)
"TF_LEN" / "TW_ZSURF"	-0.174831	(-0.201667 , -0.147995)
"TF_LEN" / "TW_PP"	-0.0176661	(-0.0656984 , 0.0303662)
"TF_LEN" / "TW_COABS"	-0.21318	(-0.238782 , -0.187578)
"TF_LEN" / "TF_COREL"	-0.255463	(-0.304734 , -0.206192)
"TF_LEN" / "TF_RG"	-0.0979306	(-0.158386 , -0.0374754)
"TF_COMP" / "TF_ZCOMB"	0.0071124	(-0.0339756 , 0.0482004)
"TF_COMP" / "TF_ZPAIR"	0.0764661	(0.034369 , 0.118563)
"TF_COMP" / "TF_ZSURF"	-0.0484198	(-0.0883656 , -0.00847409)
"TF_COMP" / "TF_PP"	0.0990678	(0.0583662 , 0.139769)
"TF_COMP" / "TF_COABS"	-0.0867421	(-0.128207 , -0.0452772)
"TF_COMP" / "TF_COREL"	-0.114293	(-0.153345 , -0.0752397)
"TF_COMP" / "TF_RG"	0.0277157	(-0.00121046 , 0.0566419)
"TF_COMP" / "TW_LEN"	0.155246	(0.108407 , 0.202086)
"TF_COMP" / "TW_COMP"	0.010438	(0.00491699 , 0.015959)
"TF_COMP" / "TW_ZCOMB"	0.0288992	(-0.0131326 , 0.0709309)
"TF_COMP" / "TW_ZPAIR"	0.0996777	(0.0571005 , 0.142255)
"TF_COMP" / "TW_ZSURF"	-0.0343488	(-0.0750603 , 0.00636275)
"TF_COMP" / "TW_PP"	0.122817	(0.0827307 , 0.162902)
"TF_COMP" / "TW_COABS"	-0.0726975	(-0.115347 , -0.0300484)
"TF_COMP" / "TF_COREL"	-0.11498	(-0.153925 , -0.0760351)
"TF_COMP" / "TF_RG"	0.0425521	(0.0140229 , 0.0710813)
"TF_ZCOMB" / "TF_ZPAIR"	0.0693537	(0.0536698 , 0.0850376)
"TF_ZCOMB" / "TF_ZSURF"	-0.0555322	(-0.073653 , -0.0374115)
"TF_ZCOMB" / "TF_PP"	0.0919554	(0.0478931 , 0.136018)
"TF_ZCOMB" / "TF_COABS"	-0.0938545	(-0.122463 , -0.0652465)
"TF_ZCOMB" / "TF_COREL"	-0.121405	(-0.165609 , -0.0772013)
"TF_ZCOMB" / "TF_RG"	0.0206033	(-0.0315831 , 0.0727898)
"TF_ZCOMB" / "TW_LEN"	0.148134	(0.121568 , 0.1747)
"TF_ZCOMB" / "TW_COMP"	0.00332562	(-0.0379492 , 0.0446004)
"TF_ZCOMB" / "TW_ZCOMB"	0.0217868	(0.00782772 , 0.0357458)
"TF_ZCOMB" / "TW_ZPAIR"	0.0925653	(0.0737164 , 0.111414)
"TF_ZCOMB" / "TW_ZSURF"	-0.0414612	(-0.0624408 , -0.0204815)
"TF_ZCOMB" / "TW_PP"	0.115704	(0.0711052 , 0.160303)
"TF_ZCOMB" / "TW_COABS"	-0.0798099	(-0.109222 , -0.0503979)
"TF_ZCOMB" / "TF_COREL"	-0.122093	(-0.166292 , -0.0778936)
"TF_ZCOMB" / "TF_RG"	0.0354397	(-0.0168875 , 0.0877669)
"TF_ZPAIR" / "TF_ZSURF"	-0.124886	(-0.154856 , -0.0949161)
"TF_ZPAIR" / "TF_PP"	0.0226017	(-0.0213849 , 0.0665882)
"TF_ZPAIR" / "TF_COABS"	-0.163208	(-0.196325 , -0.130092)
"TF_ZPAIR" / "TF_COREL"	-0.190759	(-0.236994 , -0.144523)
"TF_ZPAIR" / "TF_RG"	-0.0487504	(-0.100939 , 0.00343771)
"TF_ZPAIR" / "TW_LEN"	0.0787802	(0.0486115 , 0.108949)
"TF_ZPAIR" / "TW_COMP"	-0.0660281	(-0.108267 , -0.023789)
"TF_ZPAIR" / "TW_ZCOMB"	-0.0475669	(-0.0671468 , -0.0279871)
"TF_ZPAIR" / "TW_ZPAIR"	0.0232116	(0.00826451 , 0.0381586)
"TF_ZPAIR" / "TW_ZSURF"	-0.110815	(-0.140831 , -0.080799)
"TF_ZPAIR" / "TW_PP"	0.0463504	(0.00181547 , 0.0908854)
"TF_ZPAIR" / "TW_COABS"	-0.149164	(-0.182999 , -0.115328)
"TF_ZPAIR" / "TF_COREL"	-0.191446	(-0.237678 , -0.145214)
"TF_ZPAIR" / "TF_RG"	-0.033914	(-0.0862321 , 0.0184041)
"TF_ZSURF" / "TF_PP"	0.147488	(0.103707 , 0.191268)
"TF_ZSURF" / "TF_COABS"	-0.0383223	(-0.0634988 , -0.0131458)
"TF_ZSURF" / "TF_COREL"	-0.0658727	(-0.106216 , -0.0255295)
"TF_ZSURF" / "TF_RG"	0.0761355	(0.0261152 , 0.126156)

"TF_ZSURF" / "TW_LEN"	0.203666	(0.175495 , 0.231837)
"TF_ZSURF" / "TW_COMP"	0.0588579	(0.0185594 , 0.0991563)
"TF_ZSURF" / "TW_ZCOMB"	0.077319	(0.0552777 , 0.0993603)
"TF_ZSURF" / "TW_ZPAIR"	0.148098	(0.118325 , 0.17787)
"TF_ZSURF" / "TW_ZSURF"	0.0140711	(-0.00131322 , 0.0294554)
"TF_ZSURF" / "TW_PP"	0.171236	(0.12723 , 0.215243)
"TF_ZSURF" / "TW_COABS"	-0.0242777	(-0.0504684 , 0.00191305)
"TF_ZSURF" / "TW_COREL"	-0.0665603	(-0.106907 , -0.0262133)
"TF_ZSURF" / "TW_RG"	0.0909719	(0.0407541 , 0.14119)
"TF_PP" / "TF_COABS"	-0.18581	(-0.229515 , -0.142104)
"TF_PP" / "TF_COREL"	-0.21336	(-0.25533 , -0.171391)
"TF_PP" / "TF_RG"	-0.0713521	(-0.114148 , -0.0285563)
"TF_PP" / "TW_LEN"	0.0561785	(0.00790251 , 0.104455)
"TF_PP" / "TW_COMP"	-0.0886298	(-0.129609 , -0.0476501)
"TF_PP" / "TW_ZCOMB"	-0.0701686	(-0.114628 , -0.0257094)
"TF_PP" / "TW_ZPAIR"	0.000609917	(-0.0437575 , 0.0449774)
"TF_PP" / "TW_ZSURF"	-0.133417	(-0.176925 , -0.0899083)
"TF_PP" / "TW_PP"	0.0237488	(0.013864 , 0.0336335)
"TF_PP" / "TW_COABS"	-0.171765	(-0.216259 , -0.127271)
"TF_PP" / "TW_COREL"	-0.214048	(-0.255895 , -0.172201)
"TF_PP" / "TW_RG"	-0.0565157	(-0.0993167 , -0.0137147)
"TF_COABS" / "TF_COREL"	-0.0275504	(-0.059143 , 0.00404218)
"TF_COABS" / "TF_RG"	0.114458	(0.0643793 , 0.164536)
"TF_COABS" / "TW_LEN"	0.241988	(0.216115 , 0.267862)
"TF_COABS" / "TW_COMP"	0.0971802	(0.0554068 , 0.138954)
"TF_COABS" / "TW_ZCOMB"	0.115641	(0.0870277 , 0.144255)
"TF_COABS" / "TW_ZPAIR"	0.18642	(0.153937 , 0.218903)
"TF_COABS" / "TW_ZSURF"	0.0523934	(0.027055 , 0.0777318)
"TF_COABS" / "TW_PP"	0.209559	(0.165782 , 0.253336)
"TF_COABS" / "TW_COABS"	0.0140446	(0.00832306 , 0.0197662)
"TF_COABS" / "TW_COREL"	-0.028238	(-0.0602832 , 0.00380719)
"TF_COABS" / "TW_RG"	0.129294	(0.0790733 , 0.179515)
"TF_COREL" / "TF_RG"	0.142008	(0.108002 , 0.176014)
"TF_COREL" / "TW_LEN"	0.269539	(0.220453 , 0.318625)
"TF_COREL" / "TW_COMP"	0.124731	(0.08543 , 0.164031)
"TF_COREL" / "TW_ZCOMB"	0.143192	(0.0984829 , 0.187901)
"TF_COREL" / "TW_ZPAIR"	0.21397	(0.16751 , 0.260431)
"TF_COREL" / "TW_ZSURF"	0.0799438	(0.0390381 , 0.12085)
"TF_COREL" / "TW_PP"	0.237109	(0.19478 , 0.279438)
"TF_COREL" / "TW_COABS"	0.041595	(0.00889455 , 0.0742955)
"TF_COREL" / "TW_COREL"	-0.000687603	(-0.00551984 , 0.00414463)
"TF_COREL" / "TW_RG"	0.156845	(0.122664 , 0.191026)
"TF_RG" / "TW_LEN"	0.127531	(0.0668411 , 0.18822)
"TF_RG" / "TW_COMP"	-0.0172777	(-0.0466795 , 0.0121242)
"TF_RG" / "TW_ZCOMB"	0.00118347	(-0.0521405 , 0.0545075)
"TF_RG" / "TW_ZPAIR"	0.071962	(0.0188305 , 0.125093)
"TF_RG" / "TW_ZSURF"	-0.0620645	(-0.113188 , -0.0109411)
"TF_RG" / "TW_PP"	0.0951008	(0.0520931 , 0.138109)
"TF_RG" / "TW_COABS"	-0.100413	(-0.151645 , -0.0491817)
"TF_RG" / "TW_COREL"	-0.142696	(-0.176275 , -0.109117)
"TF_RG" / "TW_RG"	0.0148364	(0.00903575 , 0.020637)
"TW_LEN" / "TW_COMP"	-0.144808	(-0.191763 , -0.0978534)
"TW_LEN" / "TW_ZCOMB"	-0.126347	(-0.151507 , -0.101187)
"TW_LEN" / "TW_ZPAIR"	-0.0555686	(-0.0842249 , -0.0269123)
"TW_LEN" / "TW_ZSURF"	-0.189595	(-0.216195 , -0.162995)
"TW_LEN" / "TW_PP"	-0.0324298	(-0.0807582 , 0.0158987)
"TW_LEN" / "TW_COABS"	-0.227944	(-0.253274 , -0.202614)
"TW_LEN" / "TW_COREL"	-0.270226	(-0.319572 , -0.22088)
"TW_LEN" / "TW_RG"	-0.112694	(-0.173573 , -0.0518154)
"TW_COMP" / "TW_ZCOMB"	0.0184612	(-0.0234147 , 0.060337)
"TW_COMP" / "TW_ZPAIR"	0.0892397	(0.0467919 , 0.131687)
"TW_COMP" / "TW_ZSURF"	-0.0447868	(-0.0855742 , -0.00399934)
"TW_COMP" / "TW_PP"	0.112379	(0.0719322 , 0.152825)
"TW_COMP" / "TW_COABS"	-0.0831355	(-0.126001 , -0.0402705)
"TW_COMP" / "TW_COREL"	-0.125418	(-0.164764 , -0.0860727)
"TW_COMP" / "TW_RG"	0.032114	(0.00333263 , 0.0608955)
"TW_ZCOMB" / "TW_ZPAIR"	0.0707785	(0.0560483 , 0.0855087)
"TW_ZCOMB" / "TW_ZSURF"	-0.0632479	(-0.0813533 , -0.0451425)
"TW_ZCOMB" / "TW_PP"	0.0939174	(0.0491462 , 0.138689)
"TW_ZCOMB" / "TW_COABS"	-0.101597	(-0.13032 , -0.0728737)
"TW_ZCOMB" / "TW_COREL"	-0.143879	(-0.188849 , -0.0989101)
"TW_ZCOMB" / "TW_RG"	0.0136529	(-0.0398999 , 0.0672057)
"TW_ZPAIR" / "TW_ZSURF"	-0.134026	(-0.162639 , -0.105414)
"TW_ZPAIR" / "TW_PP"	0.0231388	(-0.0215308 , 0.0678085)
"TW_ZPAIR" / "TW_COABS"	-0.172375	(-0.205105 , -0.139646)
"TW_ZPAIR" / "TW_COREL"	-0.214658	(-0.261303 , -0.168013)
"TW_ZPAIR" / "TW_RG"	-0.0571256	(-0.110541 , -0.00370997)

"TW_ZSURF"/"TW_PP"	0.157165	(0.113599 , 0.200731)
"TW_ZSURF"/"TW_COABS"	-0.0383488	(-0.063899 , -0.0127985)
"TW_ZSURF"/"TW_COREL"	-0.0806314	(-0.121761 , -0.0395015)
"TW_ZSURF"/"TW_RG"	0.0769008	(0.0255684 , 0.128233)
"TW_PP"/"TW_COABS"	-0.195514	(-0.2402 , -0.150828)
"TW_PP"/"TW_COREL"	-0.237797	(-0.280032 , -0.195562)
"TW_PP"/"TW_RG"	-0.0802645	(-0.123003 , -0.0375263)
"TW_COABS"/"TW_COREL"	-0.0422826	(-0.075237 , -0.00932831)
"TW_COABS"/"TW_RG"	0.11525	(0.0638037 , 0.166695)
"TW_COREL"/"TW_RG"	0.157532	(0.123726 , 0.191339)

Anexo I: intervalos de confianza de la diferencia de las AUC para cada par de clasificadores multivariables.

TEST1/TEST2	AUC_DIFFERENCE	CONFIDENCE_INTERVAL
"nBAYES" / "BAYES_NET"	-0.0579769	(-0.0722106 , -0.0437431)
"nBAYES" / "CENTROIDS"	0.254721	(0.224506 , 0.284935)
"nBAYES" / "GA_MATH_3"	-0.033395	(-0.051018 , -0.0157721)
"nBAYES" / "GA_MATH_4"	-0.0460628	(-0.0620988 , -0.0300268)
"nBAYES" / "GA_MATH_5"	-0.0539174	(-0.0696991 , -0.0381356)
"nBAYES" / "GA_MATH_6"	-0.0539702	(-0.0707812 , -0.0371593)
"nBAYES" / "GA_LOGIC_2"	0.0495669	(0.03271 , 0.0664239)
"nBAYES" / "GA_LOGIC_3"	0.00827934	(-0.00848452 , 0.0250432)
"nBAYES" / "GA_LOGIC_4"	-0.00666281	(-0.02578 , 0.0124544)
"nBAYES" / "GA_LOGIC_5"	0.0113818	(-0.00797807 , 0.0307417)
"nBAYES" / "GA_LOGIC_6"	-0.00483967	(-0.0229368 , 0.0132574)
"nBAYES" / "GA_LOGIC_7"	-0.00783967	(-0.0255882 , 0.00990885)
"nBAYES" / "GA_LOGIC_8"	-0.00803802	(-0.0273706 , 0.0112945)
"nBAYES" / "GA_LOGIC_9"	-0.00840496	(-0.025587 , 0.00877709)
"nBAYES" / "GA_LOGIC_10"	-0.0173256	(-0.0348199 , 0.000168706)
"nBAYES" / "NN_1LAYER"	-0.0566678	(-0.0720846 , -0.0412509)
"nBAYES" / "SVM"	-0.0772562	(-0.0942244 , -0.060288)
"BAYES_NET" / "CENTROIDS"	0.312698	(0.28182 , 0.343575)
"BAYES_NET" / "GA_MATH_3"	0.0245818	(0.0116217 , 0.037542)
"BAYES_NET" / "GA_MATH_4"	0.011914	(0.0010546 , 0.0227735)
"BAYES_NET" / "GA_MATH_5"	0.0040595	(-0.00521569 , 0.0133347)
"BAYES_NET" / "GA_MATH_6"	0.00400661	(-0.00604489 , 0.0140581)
"BAYES_NET" / "GA_LOGIC_2"	0.107544	(0.087744 , 0.127344)
"BAYES_NET" / "GA_LOGIC_3"	0.0662562	(0.0518667 , 0.0806457)
"BAYES_NET" / "GA_LOGIC_4"	0.051314	(0.0360678 , 0.0665603)
"BAYES_NET" / "GA_LOGIC_5"	0.0693587	(0.0529994 , 0.0857179)
"BAYES_NET" / "GA_LOGIC_6"	0.0531372	(0.0379583 , 0.0683161)
"BAYES_NET" / "GA_LOGIC_7"	0.0501372	(0.034301 , 0.0659734)
"BAYES_NET" / "GA_LOGIC_8"	0.0499388	(0.034762 , 0.0651157)
"BAYES_NET" / "GA_LOGIC_9"	0.0495719	(0.0349924 , 0.0641515)
"BAYES_NET" / "GA_LOGIC_10"	0.0406512	(0.026985 , 0.0543175)
"BAYES_NET" / "NN_1LAYER"	0.00130909	(-0.00784828 , 0.0104665)
"BAYES_NET" / "SVM"	-0.0192793	(-0.028595 , -0.00996373)
"CENTROIDS" / "GA_MATH_3"	-0.288116	(-0.320529 , -0.255703)
"CENTROIDS" / "GA_MATH_4"	-0.300783	(-0.331664 , -0.269903)
"CENTROIDS" / "GA_MATH_5"	-0.308638	(-0.340376 , -0.2769)
"CENTROIDS" / "GA_MATH_6"	-0.308691	(-0.340948 , -0.276434)
"CENTROIDS" / "GA_LOGIC_2"	-0.205154	(-0.238195 , -0.172113)
"CENTROIDS" / "GA_LOGIC_3"	-0.246441	(-0.278861 , -0.214021)
"CENTROIDS" / "GA_LOGIC_4"	-0.261383	(-0.295476 , -0.227291)
"CENTROIDS" / "GA_LOGIC_5"	-0.243339	(-0.277179 , -0.209499)
"CENTROIDS" / "GA_LOGIC_6"	-0.25956	(-0.292225 , -0.226896)
"CENTROIDS" / "GA_LOGIC_7"	-0.26256	(-0.293483 , -0.231638)
"CENTROIDS" / "GA_LOGIC_8"	-0.262759	(-0.29617 , -0.229347)
"CENTROIDS" / "GA_LOGIC_9"	-0.263126	(-0.294774 , -0.231478)
"CENTROIDS" / "GA_LOGIC_10"	-0.272046	(-0.303775 , -0.240318)
"CENTROIDS" / "NN_1LAYER"	-0.311388	(-0.34286 , -0.279917)
"CENTROIDS" / "SVM"	-0.331977	(-0.364096 , -0.299857)
"GA_MATH_3" / "GA_MATH_4"	-0.0126678	(-0.0213484 , -0.00398718)
"GA_MATH_3" / "GA_MATH_5"	-0.0205223	(-0.0321946 , -0.00884999)
"GA_MATH_3" / "GA_MATH_6"	-0.0205752	(-0.0306685 , -0.0104819)
"GA_MATH_3" / "GA_LOGIC_2"	0.082962	(0.0643652 , 0.101559)
"GA_MATH_3" / "GA_LOGIC_3"	0.0416744	(0.0257332 , 0.0576156)
"GA_MATH_3" / "GA_LOGIC_4"	0.0267322	(0.0100308 , 0.0434336)
"GA_MATH_3" / "GA_LOGIC_5"	0.0447769	(0.027451 , 0.0621028)
"GA_MATH_3" / "GA_LOGIC_6"	0.0285554	(0.0122032 , 0.0449075)
"GA_MATH_3" / "GA_LOGIC_7"	0.0255554	(0.00868579 , 0.042425)
"GA_MATH_3" / "GA_LOGIC_8"	0.025357	(0.00916304 , 0.041551)
"GA_MATH_3" / "GA_LOGIC_9"	0.0249901	(0.00956897 , 0.0404112)
"GA_MATH_3" / "GA_LOGIC_10"	0.0160694	(0.00013096 , 0.0320079)
"GA_MATH_3" / "NN_1LAYER"	-0.0232727	(-0.032678 , -0.0138675)
"GA_MATH_3" / "SVM"	-0.0438612	(-0.05718 , -0.0305423)
"GA_MATH_4" / "GA_MATH_5"	-0.00785455	(-0.0148772 , -0.000831898)
"GA_MATH_4" / "GA_MATH_6"	-0.00790744	(-0.0134736 , -0.00234124)
"GA_MATH_4" / "GA_LOGIC_2"	0.0956298	(0.0761915 , 0.115068)
"GA_MATH_4" / "GA_LOGIC_3"	0.0543421	(0.0386594 , 0.0700249)
"GA_MATH_4" / "GA_LOGIC_4"	0.0394	(0.0236904 , 0.0551096)
"GA_MATH_4" / "GA_LOGIC_5"	0.0574446	(0.0400468 , 0.0748424)
"GA_MATH_4" / "GA_LOGIC_6"	0.0412231	(0.0249952 , 0.0574511)
"GA_MATH_4" / "GA_LOGIC_7"	0.0382231	(0.02264 , 0.0538063)

"GA_MATH_4"/"GA_LOGIC_8"	0.0380248	(0.0220135 , 0.0540361)
"GA_MATH_4"/"GA_LOGIC_9"	0.0376579	(0.0230091 , 0.0523066)
"GA_MATH_4"/"GA_LOGIC_10"	0.0287372	(0.0143834 , 0.043091)
"GA_MATH_4"/"NN_1LAYER"	-0.010605	(-0.0177688 , -0.00344108)
"GA_MATH_4"/"SVM"	-0.0311934	(-0.0431746 , -0.0192122)
"GA_MATH_5"/"GA_MATH_6"	-5.28926e-05	(-0.00492665 , 0.00482087)
"GA_MATH_5"/"GA_LOGIC_2"	0.103484	(0.0831803 , 0.123788)
"GA_MATH_5"/"GA_LOGIC_3"	0.0621967	(0.0464428 , 0.0779506)
"GA_MATH_5"/"GA_LOGIC_4"	0.0472545	(0.0319555 , 0.0625536)
"GA_MATH_5"/"GA_LOGIC_5"	0.0652992	(0.0479057 , 0.0826927)
"GA_MATH_5"/"GA_LOGIC_6"	0.0490777	(0.0327452 , 0.0654102)
"GA_MATH_5"/"GA_LOGIC_7"	0.0460777	(0.0298613 , 0.0622941)
"GA_MATH_5"/"GA_LOGIC_8"	0.0458793	(0.0298639 , 0.0618948)
"GA_MATH_5"/"GA_LOGIC_9"	0.0455124	(0.0306832 , 0.0603416)
"GA_MATH_5"/"GA_LOGIC_10"	0.0365917	(0.0221955 , 0.050988)
"GA_MATH_5"/"NN_1LAYER"	-0.00275041	(-0.0100538 , 0.004553)
"GA_MATH_5"/"SVM"	-0.0233388	(-0.0336837 , -0.012994)
"GA_MATH_6"/"GA_LOGIC_2"	0.103537	(0.0833244 , 0.12375)
"GA_MATH_6"/"GA_LOGIC_3"	0.0622496	(0.0464151 , 0.0780841)
"GA_MATH_6"/"GA_LOGIC_4"	0.0473074	(0.0320058 , 0.0626091)
"GA_MATH_6"/"GA_LOGIC_5"	0.0653521	(0.0476427 , 0.0830615)
"GA_MATH_6"/"GA_LOGIC_6"	0.0491306	(0.0324628 , 0.0657983)
"GA_MATH_6"/"GA_LOGIC_7"	0.0461306	(0.0295566 , 0.0627046)
"GA_MATH_6"/"GA_LOGIC_8"	0.0459322	(0.0296748 , 0.0621897)
"GA_MATH_6"/"GA_LOGIC_9"	0.0455653	(0.0304214 , 0.0607092)
"GA_MATH_6"/"GA_LOGIC_10"	0.0366446	(0.021957 , 0.0513323)
"GA_MATH_6"/"NN_1LAYER"	-0.00269752	(-0.00983763 , 0.00444259)
"GA_MATH_6"/"SVM"	-0.023286	(-0.0340066 , -0.0125653)
"GA_LOGIC_2"/"GA_LOGIC_3"	-0.0412876	(-0.0579568 , -0.0246184)
"GA_LOGIC_2"/"GA_LOGIC_4"	-0.0562298	(-0.0769896 , -0.0354699)
"GA_LOGIC_2"/"GA_LOGIC_5"	-0.0381851	(-0.0571633 , -0.0192069)
"GA_LOGIC_2"/"GA_LOGIC_6"	-0.0544066	(-0.0720518 , -0.0367614)
"GA_LOGIC_2"/"GA_LOGIC_7"	-0.0574066	(-0.0743665 , -0.0404467)
"GA_LOGIC_2"/"GA_LOGIC_8"	-0.057605	(-0.0768032 , -0.0384068)
"GA_LOGIC_2"/"GA_LOGIC_9"	-0.0579719	(-0.0742463 , -0.0416975)
"GA_LOGIC_2"/"GA_LOGIC_10"	-0.0668926	(-0.0853987 , -0.0483864)
"GA_LOGIC_2"/"NN_1LAYER"	-0.106235	(-0.124742 , -0.087727)
"GA_LOGIC_2"/"SVM"	-0.126823	(-0.14693 , -0.106716)
"GA_LOGIC_3"/"GA_LOGIC_4"	-0.0149421	(-0.0308425 , 0.000958218)
"GA_LOGIC_3"/"GA_LOGIC_5"	0.00310248	(-0.0122749 , 0.0184798)
"GA_LOGIC_3"/"GA_LOGIC_6"	-0.013119	(-0.0270157 , 0.000777714)
"GA_LOGIC_3"/"GA_LOGIC_7"	-0.016119	(-0.0331047 , 0.000866656)
"GA_LOGIC_3"/"GA_LOGIC_8"	-0.0163174	(-0.031097 , -0.00153776)
"GA_LOGIC_3"/"GA_LOGIC_9"	-0.0166843	(-0.0313701 , -0.00199847)
"GA_LOGIC_3"/"GA_LOGIC_10"	-0.025605	(-0.0375026 , -0.0137073)
"GA_LOGIC_3"/"NN_1LAYER"	-0.0649471	(-0.079409 , -0.0504852)
"GA_LOGIC_3"/"SVM"	-0.0855355	(-0.101597 , -0.0694744)
"GA_LOGIC_4"/"GA_LOGIC_5"	0.0180446	(0.00400486 , 0.0320844)
"GA_LOGIC_4"/"GA_LOGIC_6"	0.00182314	(-0.0139521 , 0.0175984)
"GA_LOGIC_4"/"GA_LOGIC_7"	-0.00117686	(-0.0196669 , 0.0173131)
"GA_LOGIC_4"/"GA_LOGIC_8"	-0.00137521	(-0.0146176 , 0.0118672)
"GA_LOGIC_4"/"GA_LOGIC_9"	-0.00174215	(-0.0176378 , 0.0141535)
"GA_LOGIC_4"/"GA_LOGIC_10"	-0.0106628	(-0.0266297 , 0.00530404)
"GA_LOGIC_4"/"NN_1LAYER"	-0.050005	(-0.0652008 , -0.0348091)
"GA_LOGIC_4"/"SVM"	-0.0705934	(-0.0868182 , -0.0543686)
"GA_LOGIC_5"/"GA_LOGIC_6"	-0.0162215	(-0.0278502 , -0.00459276)
"GA_LOGIC_5"/"GA_LOGIC_7"	-0.0192215	(-0.036437 , -0.00200596)
"GA_LOGIC_5"/"GA_LOGIC_8"	-0.0194198	(-0.0300911 , -0.00874861)
"GA_LOGIC_5"/"GA_LOGIC_9"	-0.0197868	(-0.0353911 , -0.00418244)
"GA_LOGIC_5"/"GA_LOGIC_10"	-0.0287074	(-0.0428976 , -0.0145173)
"GA_LOGIC_5"/"NN_1LAYER"	-0.0680496	(-0.0841406 , -0.0519586)
"GA_LOGIC_5"/"SVM"	-0.088638	(-0.106278 , -0.0709983)
"GA_LOGIC_6"/"GA_LOGIC_7"	-0.003	(-0.0181849 , 0.0121849)
"GA_LOGIC_6"/"GA_LOGIC_8"	-0.00319835	(-0.0136664 , 0.00726967)
"GA_LOGIC_6"/"GA_LOGIC_9"	-0.00356529	(-0.0157875 , 0.00865691)
"GA_LOGIC_6"/"GA_LOGIC_10"	-0.012486	(-0.0249646 , -7.33605e-06)
"GA_LOGIC_6"/"NN_1LAYER"	-0.0518281	(-0.0667349 , -0.0369213)
"GA_LOGIC_6"/"SVM"	-0.0724165	(-0.0887457 , -0.0560874)
"GA_LOGIC_7"/"GA_LOGIC_8"	-0.000198347	(-0.0171268 , 0.0167301)
"GA_LOGIC_7"/"GA_LOGIC_9"	-0.000565289	(-0.0142802 , 0.0131497)
"GA_LOGIC_7"/"GA_LOGIC_10"	-0.00948595	(-0.0243247 , 0.00535277)
"GA_LOGIC_7"/"NN_1LAYER"	-0.0488281	(-0.064336 , -0.0333202)
"GA_LOGIC_7"/"SVM"	-0.0694165	(-0.0854542 , -0.0533789)
"GA_LOGIC_8"/"GA_LOGIC_9"	-0.000366942	(-0.014276 , 0.0135421)
"GA_LOGIC_8"/"GA_LOGIC_10"	-0.0092876	(-0.0233575 , 0.00478234)
"GA_LOGIC_8"/"NN_1LAYER"	-0.0486298	(-0.063431 , -0.0338285)
"GA_LOGIC_8"/"SVM"	-0.0692182	(-0.085352 , -0.0530843)

"GA_LOGIC_9"/"GA_LOGIC_10"	-0.00892066	(-0.0214928 , 0.0036515)
"GA_LOGIC_9"/"NN_1LAYER"	-0.0482628	(-0.0622207 , -0.0343049)
"GA_LOGIC_9"/"SVM"	-0.0688512	(-0.0848901 , -0.0528124)
"GA_LOGIC_10"/"NN_1LAYER"	-0.0393421	(-0.0532376 , -0.0254467)
"GA_LOGIC_10"/"SVM"	-0.0599306	(-0.0745573 , -0.0453039)
"NN_1LAYER"/"SVM"	-0.0205884	(-0.0298726 , -0.0113043)