Análisis del virus de la gripe A mediante descubrimientos de subgrupos difusos

C.J. Carmona, C. Chrysostomou, H. Seker, M.J. del Jesus

Department of Computer Science, University of Jaen, Spain [1]
{ccarmona,mjjesus}@ujaen.es
Department of Genetics, University of Leicester, Leicester, United Kingdom [2]
cc390@le.ac.uk
Centre for Computational Intelligence, De Montfort University, Leicester, United Kingdom [3]
hseker@dmu.ac.uk

Resumen El virus de la gripe A está siendo en los últimos años uno de los principales problemas de pandemias mortales de los Siglos XX y XXI. Las diferentes mutaciones que sufre este virus desembocan en una gran dificultad para la creación de vacunas y/o medicinas que puedan combatirlo.
En este estudio experimental se buscan relaciones interesantes y atípicas entre diferentes proteínas del virus de la gripe A. Con estas relaciones, se buscan las propiedades capaces de distinguir y describir los distintos tipos de virus que podrían proporcionar a los expertos información que ayude en el desarrollo de nuevas terapias para este virus. Para ello, se estudian un conjunto de proteínas recogidas en los últimos años. Para cumplir este objetivo, se realizan inicialmente unas transformaciones de las cadenas proteicas del virus para su posterior análisis con el algoritmo de descubrimiento de subgrupos basado en sistemas difusos evolutivos más destacado, el algoritmo NMEEF-SD.

Keywords: Sistemas difusos evolutivos, Virus de la gripe A, Descubrimiento de subgrupos

1. Introducción

El virus de la gripe A pertenece a la familia *Orthomyxoviridae* y afecta principalmente en aves y algunos mamíferos. El genoma de este virus está formado por 8 genes sencillos: el gen *hemagglutinin* (HA), el gen *neuraminidase* (NA), el gen *nucleoprotein* (NP), el gen *matrix proteins* (M), el gen *non-structural proteins* (NS) y los tres genes *RNA polymerase* (PA, PB1, PB2). Raras veces surgen brotes o pandemias cuando el virus de la gripe A se transmite de aves salvajes a aves domésticas.

Durante el Siglo XX se han registrado tres grandes pandemias provocadas por el virus de la gripe A dentro de la raza humana, concretamente causadas por los subtipos de virus H1N1, H2N2 y H3N2. Además de estos tres subtipos, dentro del

virus de la gripe A, el H5N1 se considera como conductor de la pandemia actual. En este análisis, se utilizan estos cuatro subtipos del virus de la gripe A, que son el objetivo principal de estudio para la creación de medicinas o antivirales, que se denominan inhibidores de NA [13].

A lo largo de los años se ha recogido información referente a estos subtipos de virus [2]: para el subtipo H1N1 se han recogido 200 proteínas desde el 2009, para el H2N2 se han recogido 76 entre los años 1957 y 1968; para el subtipo H3N2 se han recogido 200 desde el periodo 1968 hasta el 2000 y para el subtipo H5N1 se han recogido 70 proteínas desde 2005 a 2009. La relación de estos subtipos del virus de la gripe A con respecto al gen NA es la siguiente: El virus H1N1 es el resultado de reordenaciones entre el virus H1N1 euro-asiático del cerdo y el virus H1N2 del cerdo, el virus H2N2 es el resultado de la reordenación entre el virus H1N1 humano y el virus de la gripe aviar H2N2, el virus H3N2 es el resultado de la reordenación entre el virus H2N2 circulante entre humanos y el virus de las aves H3 y el virus H5N1 fue creado mediante diversas combinaciones de subtipos de virus de la gripe A.

Para el análisis del problema, este trabajo se centra en la técnica de minería de datos del Descubrimiento de Subgrupos (SD) [10] cuyo principal objetivo es la obtención de relaciones parciales en los datos con estadísticas inusuales y de interés con respecto a una variable objetivo. Para ello, se va a aplicar el algoritmo NMEEF-SD [3] que es en la actualidad el algoritmo de SD basado en sistemas difusos evolutivos (EFSs) [9] más destacado de la literatura. Los EFSs están basados en lógica difusa y permiten trabajar en entornos con variables continuas sin necesidad de una previa discretización como es el problema que se presenta en este trabajo.

Este trabajo se divide en las siguientes secciones: En la Sección 2, se puede observar la transformación llevada a cabo sobre las proteinas para prepararlas y poder aplicar SD, en la Sección 3 se presenta de forma general SD y las ventajas de aplicar el algoritmo NMEEF-SD y en la Sección 4 se presenta el estudio experimental realizado. Para finalizar se presentan las conclusiones obtenidas en el trabajo.

2. Procesamiento de señal para el análisis de secuencias de proteínas

Recientemente, se han utilizado diversos métodos dentro de la bioinformática para el análisis de secuencias de proteínas, donde algunos de los más comunes son el *Resonant Recognition Model* [5,6] y el *Complex Resonant Recognition Model* [4]. Estudios previos [15] han utilizado los subtipos del virus de la gripe A para analizar el gen HA con el *Resonant Recognition Model* con el objetivo de identificar nuevas terapias que permitan el desarrollo de nuevas medicinas así como comprender cómo interacciona el virus de la gripe con sus receptores.

A diferencia de otros estudios previos, este estudio experimental ha sido realizado directamente mediante un espectro absoluto que se deriva de la aplicación

de la transformación discreta de fourier para cada secuencia proteica con codificación numérica. Para poder aplicar la función de fourier, es necesario utilizar un índice de aminoácido, como el *electron-ion interaction potential* (EIIP) [14]. Mediante este índice, mostrado en la tabla 1, se convierten las secuencias proteicas en secuencias numéricas.

Tabla 1. Valores del índice *electron-ion interaction potential*

Amino	***EIIP***	***Amino***	***EIIP***	***Amino***	***EIIP***	***Amino***	***EIIP***
Leu	0.0000	Tyr	0.0516	Ile	0.0000	Trp	0.0548
Asn	0.0036	Gln	0.0761	Gly	0.0050	Met	0.0823
Glu	0.0057	Ser	0.0829	Val	0.0058	Cys	0.0829
Pro	0.0198	Thr	0.0941	His	0.0242	Phe	0.0946
Lys	0.0371	Arg	0.0959	Ala	0.0373	Asp	0.1263

La transformación discreta de fourier se define mediante la ecuación 1:

$$X(n) = \sum_{m=0}^{N-1} x(m) e^{-j(2/N)nm} \quad n = 1, 2, \ldots, N/2 \tag{1}$$

donde $x(m)$ es el valor de la posición m de la serie numérica, N es el número de puntos en la serie, y $X(n)$ son los coeficientes de la transformada. La máxima frecuencia del espectro viene determinada por la siguiente ecuación:

$$F = \frac{1}{2d} \tag{2}$$

donde F es la frecuencia máxima y d es la distancia entre puntos de la secuencia.

Si se asume que todos los puntos de la secuencia son equidistantes con una distancia $d = 1$ entonces la frecuencia máxima del espectro sería $F = \frac{1}{2(1)} = 0{,}5$. Esto indica que el rango de frecuencia no depende del número de puntos en la secuencia sino de la resolución del espectro. La salida de la transformada de Fourier es una secuencia que se puede representar como indica la ecuación 3.

$$X(n) = (R(n) + I(n)j), \quad n = 1, 2, \ldots, N/2 \tag{3}$$

donde $R(n)$ es la parte real de la secuencia y la función $I(n)j$ la parte imaginaria.

El paso final en el cálculo del espectro absoluto de la transformada se calcula mediante la ecuación 4.

$$S_a(n) = X(n)X*(n) = |X(n)|^2, \quad n = 1, 2, \ldots, N/2 \tag{4}$$

donde S_a es el espectro absoluto para una proteína específica, $X(n)$ son los coeficientes de la transformada en las series de $x(n)$ y $X*(n)$ son los complejos conjugados. Para escalar el espectro absoluto se utiliza la ecuación 5.

$$V = \frac{\sqrt{\sum_{n=0}^{L} C_a(n)}}{L} \tag{5}$$

donde L es el número de puntos en el espectro absoluto (C_a).

Para el análisis de las proteínas del virus de la gripe A, como las secuencias tienen diferentes longitudes, la técnica del relleno vacío (completar con 0 las variables vacías) se emplea para extender las secuencias hasta un valor de $N = 512$, de forma que la salida del espectro absoluto (ec. 4) tenga 256 propiedades.

3. Descubrimiento de subgrupos y su aplicación al problema del virus de la gripe A

En esta sección se describe brevemente por un lado la técnica de SD y por otro las ventajas proporcionadas por el algoritmo NMEEF-SD en este problema.

3.1. Descubrimiento de subgrupos

El SD es un tipo de inducción descriptiva que pretende generar modelos basados en reglas cuya finalidad es descriptiva, empleando una perspectiva predictiva para obtenerlos [11,16]. Se trata por tanto de una tarea con objetivos básicamente descriptivos que incluye características de la inducción predictiva. Este concepto se define como [17]:

> *En SD, asumimos una población de individuos dada (objetos, clientes, ...) y una propiedad de estos individuos en la que estemos interesados. La tarea del SD es entonces descubrir los subgrupos de la población que son estadísticamente "más interesantes", es decir, individuos que sean tan grandes como sea posible y tenga una distribución estadística los más atípica posible, con respecto a una propiedad de interés.*

Así, una regla (R), que consiste de una descripción de un subgrupo inducido, puede ser definida formalmente como [12]:

$$R : Cond \rightarrow VarObj$$

donde $VarObj$ es el valor de la variable de interés o variable objetivo para la tarea de SD (puede aparecer además en la bibliografía específica como $Clase$), y $Cond$ es comúnmente una conjunción de funciones (pares atributo-valor) que es capaz de describir una distribución estadística inusual con respecto a la variable objetivo.

Existen diferentes elementos a especificar en el diseño de un algoritmo de SD [1], donde uno de los más destacados son las medidas de calidad utilizadas para el proceso de búsqueda y/o evaluación de los algoritmos. A continuación, se detallan las medidas de calidad más utilizadas en la literatura y en este trabajo:

- *Atipicidad*: Esta medida se describe como el balance entre la cobertura de la regla y la ganancia de precisión [12]. Se puede calcular como:

$$Atip(R) = \frac{n(Cond)}{n_s}\left(\frac{n(VarObj \cdot Cond)}{n(Cond)} - \frac{n(VarObj)}{n_s}\right) \qquad (6)$$

donde n_s es el número de ejemplos, $n(Cond)$ es el número de ejemplos que satisfacen la condición de la regla, $n(VarObj \cdot Cond)$ es el número de ejemplos que satisfacen la condición y además pertenecen al valor de la variable objetivo en la regla y $n(VarObj)$ son todos los ejemplos del valor de la variable objetivo.

- *Sensibilidad*: Esta medida mide la proporción de ejemplos correctamente descritos [11]. Se puede calcular como:

$$Sens(R) = \frac{n(VarObj \cdot Cond)}{n(VarObj)} \tag{7}$$

Esta medida de calidad se utiliza ara evaluar la calidad de los subgrupos en el espacio ROC (*Receiver Operating Characteristic*). La medida de sensibilidad combina la precisión y generalidad generada para un valor de la variable objetivo.

- *Confianza difusa*: Determina la frecuencia relativa de los ejemplos que satisfacen tanto el antecedente como el consecuente de una regla entre aquellos que satisfacen sólo el antecedente [7]. Se calcula como:

$$CnfD(R) = \frac{\sum\limits_{E^k \in E / E^k \in VarObj} APC(E^k, R)}{\sum\limits_{E^k \in E} APC(E^k, R)} \tag{8}$$

donde APC es el grado de compatibilidad entre un ejemplo y el antecedente de una regla difusa.

3.2. Aplicación al problema del virus de la gripe A del algoritmo NMEEF-SD

Tradicionalmente, el problema del virus de la gripe A se ha resuelto utilizando clasificadores. Sin embargo, el principal inconveniente de la utilización de los clasificadores para resolver problemas de bioinformática es, en general, la falta de interpretabilidad obtenida por los modelos. Esto se debe a que los modelos extraídos tienen la exactitud como principal objetivo, lo que provoca la obtención de modelos de una cierta complejidad, ya que utilizan un amplio número de variables o propiedades para describir diferentes virus del conjunto de datos. De esta forma, es muy difícil para los expertos analizar y comprender el comportamiento de un conjunto de datos con respecto a una variable de interés. Por el contrario, los algoritmos de SD extraen modelos sencillos, con pocas reglas y un bajo número de variables, para una variable objetivo.

La búsqueda de reglas interesantes y atípicas por los algoritmos de SD es una de las ventajas proporcionadas por la aplicación del algoritmo NMEEF-SD. Para este problema, el algoritmo utiliza las medidas de atipicidad (ec. 6) y sensibilidad (ec. 7) como vectores objetivo del enfoque multi-objetivo permitiendo además maximizar, no solo estas medidas, sino también otras medidas de la tarea de SD como la confianza.

Otra de las ventajas de la aplicación de NMEEF-SD es la utilización de lógica difusa [18] para resolver el problema, ya que la obtención de reglas difusas facilita el análisis a los expertos porque se emplean etiquetas lingüísticas en todas las variables del conjunto de datos, lo que proporciona a los expertos un conocimiento más cercano al razonamiento humano, empleando valores del lenguaje natural en vez de intervalos numéricos.

Por todo ello, NMEEF-SD es un algoritmo basado en un sistema multiobjetivo difuso evolutivo [8] que contribuye a extraer conocimiento novedoso y relevante sobre relaciones entre las propiedades del problema y diferentes tipos del virus de la gripe A.

4. Estudio experimental

El problema tiene una alta dimensionalidad y está compuesto por 256 variables y 546 secuencias de proteínas distribuidas de la siguiente forma: 200 secuencias del subtipo H1N1, 76 del H2N2, 200 del H3N2 y 70 del subtipo H5N1. Todas las variables son continuas y toman valores en el dominio de los números reales. El algoritmo NMEEF-SD considera las variables continuas como variables difusas lingüísticas aplicando lógica difusa. Más concretamente, en este problema se emplean funciones de pertenencia triangulares para las variables.

Los parámetros utilizados por el algoritmo NMEEF-SD son: tamaño de la población=50, evaluaciones=10000, probabilidad de cruce=0.6, mutación=0.1, etiquetas={3,5,7,9}, objetivos={atipicidad, sensibilidad} y confianza mínima={0.2, 0.4, 0.6}

Debido a la naturaleza no determinística del algoritmo NMEEF-SD, se ha aplicado un esquema de validación cruzada de 5 particiones, con 5 ejecuciones por partición. De esta forma, los resultados que se muestran son la media de los resultados obtenidos para cada conjunto de datos para las diferentes ejecuciones, es decir la media de las 25 ejecuciones (5 particiones x 5 ejecuciones de cada partición). En cada tabla, se muestran los valores de: número de etiquetas lingüísticas, umbral mínimo de confianza empleado (Min_{Cnf}), número de reglas (n_r), número de variables (n_v), atipicidad ($ATIP$), sensibilidad ($SENS$) y confianza ($CONF$).

El estudio experimental que se presenta a continuación consta de dos partes. Por una parte, en la sección 4.1 se estudian los resultados de la aplicación del algoritmo NMEEF-SD mientras por otro lado, en la sección 4.2 se aplica NMEEF-SD al conjunto de datos completo para obtener información descriptiva acerca de los diferentes tipos de virus estudiados en el problema.

4.1. Análisis de los resultados obtenidos por el algoritmo NMEEF-SD

Debido a la complejidad del problema se han utilizado diferentes número de etiquetas lingüísticas por variable y distintos umbrales de confianza mínima para

Tabla 2. Resultados obtenidos por el algoritmo NMEEF-SD

ELs	Min_{Cnf}	n_r	n_v	*ATIP*	*SENS*	*CONF*
3	0.2	4.60	2.79	0.153	1.000	0.747
	0.4	3.80	2.65	0.174	1.000	0.811
	0.6	2.60	2.73	0.190	1.000	0.849
5	0.2	3.40	2.13	0.125	0.990	0.708
	0.4	3.00	2.17	0.134	0.992	0.767
	0.6	2.20	2.10	0.148	1.000	0.807
7	0.2	3.00	2.28	0.110	0.963	0.760
	0.4	2.40	2.42	0.113	0.939	0.854
	0.6	1.60	2.37	0.127	0.938	0.911
9	0.2	1.60	2.00	0.092	0.952	0.585
	0.4	1.40	2.00	0.099	0.944	0.631
	0.6	0.60	0.80	0.048	0.378	0.394

encontrar la configuración del algoritmo que obtenga los mejores resultados para el mismo. Los resultados se muestran en la tabla 2.

En general, se puede observar que los mejores resultados se obtienen con el uso de 3 etiquetas lingüísticas y con un umbral de confianza de 0.6. Sin embargo, el número de reglas obtenido es inferior al número de virus analizados en el conjunto de datos, lo que indica que el algoritmo no ha obtenido reglas para describir todos los subtipos de virus. Por ello, se debe realizar un análisis de los subgrupos obtenidos por el algoritmo con 3 etiquetas lingüísticas y poder establecer la mejor configuración del algoritmo a este problema. Los resultados de este análisis se presentan en la tabla 3, donde se muestran los resultados de todos los subgrupos obtenidos en cada grupo de la validación cruzada para cada subtipo de virus.

Tabla 3. Resultados para cada subtipo de virus con 3 etiquetas lingüísticas

Min_{Cnf}	*Virus*	n_r	n_v	*ATIP*	*SENS*	*CONF*
0.2	H1N1	8.00	2.88	0.199	1.000	0.849
	H2N2	5.00	3.20	0.101	1.000	0.543
	H3N2	6.00	2.50	0.178	1.000	0.812
	H5N1	5.00	2.60	0.102	1.000	0.717
0.4	H1N1	8.00	2.88	0.199	1.000	0.849
	H2N2	3.00	2.33	0.107	1.000	0.601
	H3N2	5.00	2.40	0.193	1.000	0.835
	H5N1	3.00	3.00	0.104	1.000	0.768
0.6	H1N1	7.00	3.00	0.202	1.000	0.867
	H2N2	0.00	0.00	0.000	0.000	0.000
	H3N2	5.00	2.40	0.193	1.000	0.835
	H5N1	1.00	3.00	0.101	1.000	0.867

Como se ha mencionado anteriormente en el análisis de la tabla 2 y con los resultados mostrados en la tabla 3, los subgrupos obtenidos para un umbral de confianza de 0.6 indica que no hay subgrupos para poder describir todos los subtipos de virus. Esto se debe a que el nivel de confianza es muy alto para obtener buenos resultados en todas las virus. Por ello, los resultados obtenidos en esta configuración deben ser descartados.

Por tanto, los mejores resultados para el algoritmo NMEEF-SD se obtienen con 3 etiquetas lingüísticas y un umbral de confianza mínimo de 0.2 y 0.4. Este estudio se completa con un análisis de los subgrupos obtenidos para cada virus:

- Los subgrupos obtenidos para el virus H1N1 tienen una alta interpretabilidad porque el número de variables es bajo, donde en general los subgrupos obtenidos tienen menos de 3 variables (considerando también la variable objetivo como una variable). Los valores para la medida de atipicidad son los más altos con respecto a los valores obtenidos en el resto de clases. Además, la relación entre sensibilidad y confianza es muy bueno, ya que el algoritmo obtiene subgrupos donde todas las secuencias de proteínas para los virus son cubiertas y la confianza está cercana al 85 %.
- Para el virus H2N2 se obtienen los subgrupos con el menor número de variables por lo que la interpretabilidad en este virus es excelente. Los valores de atipicidad son además altos considerando que este virus tiene un número muy bajo de secuencias en el conjunto de datos. El nivel de sensibilidad de los subgrupos extraídos es el máximo y el valor de la confianza es bueno ya que los subgrupos superan el 60 %.
- En el virus H3N2 se obtienen los mejores subgrupos juntos con el virus H1N1, donde la interpretabilidad y los valores de la atipicidad, sensibilidad y confianza son muy altos.
- El virus H5N1 es el subtipo con menor número de proteínas. A pesar de esto, los resultados de sensibilidad y confianza son muy interesantes porque los subgrupos cubren todos los ejemplos del virus con un buen nivel de confianza (más del 70 %). Los resultados para la relevancia y atipicidad son además aceptables.

4.2. Subgrupos difusos extraídos por el algoritmo NMEEF-SD

Una vez analizado en la etapa anterior la configuración con mejores resultados para el algoritmo, a continuación se realiza un nuevo experimento utilizando el conjunto de datos completo, para analizar los subgrupos obtenidos por el algoritmo que puedan ser de interés para los expertos con un umbral mínimo de confianza de 0.2 y 3 etiquetas lingüísticas.

La tabla 4 muestra los subgrupos obtenidos por el algoritmo NMEEF-SD para cada clase, en la que la propiedad número x se identifica con el nombre $f(x)$. La tabla presenta además los resultados asociados a cada subgrupo.

Tabla 4. Subgrupos obtenidos por el algoritmo NMEEF-SD

Subgrupo	***ATIP***	***SENS***	***CONF***
$SI\ (f44 = Bajo\ Y\ f97 = Bajo) \rightarrow$ H1N1	0.224	1.000	0.966
$SI\ (f9 = Bajo\ Y\ f54 = Bajo\ f153 = Bajo\ Y\ f217 = Bajo) \rightarrow$ H2N2	0.105	1.000	0.600
$SI\ (f8 = Bajo) \rightarrow$ H3N2	0.182	1.000	0.730
$SI\ (f141 = Bajo\ Y\ f207 = Bajo\ Y\ f219 = Bajo) \rightarrow$ H3N2	0.196	0.995	0.966
$SI\ (f115 = Bajo) \rightarrow$ H5N1	0.097	1.000	0.677

Como se puede observar en la tabla 4, los buenos resultados en atipicidad muestran conocimiento novedoso y desconocido del problema. Además, la sensibilidad obtenida para la mayoría de los subgrupos tiene el máximo nivel (100 %) y la confianza es muy alta con valores que están por encima del 60 % y algunos muy cercanos al máximo nivel. Estas buenas relaciones entre los valores de sensibilidad y confianza presentan subgrupos de alta calidad. La interpretabilidad es también excelente, obteniendo subgrupos que en ningún caso superan las 4 variables, es decir se obtienen subgrupos con únicamente 4 variables como máximo de un total de 256.

Otros métodos que utilizan técnicas de procesamiento para extraer propiedades biológicamente relacionadas para caracterizar secuencias de proteínas, como el *Resonant Recognition Model* en el gen HA [15] y *Complex Resonant Recognition* para el gen NA [4], emplean análisis informativos de espectro para caracterizar un tipo de virus específico o compararlo con otras proteínas basadas en picos de frecuencia comunes [4]. Mediante el uso del algoritmo NMEEF-SD, tal y como se muestra en la tabla 4, se pueden extraer reglas sencillas basadas en la recuperación de propiedades del espectro absoluto, con respecto al virus de la gripe A. Con estas propiedades se puede obtener conocimiento que permita mejorar el análisis sobre este tipo de virus, ya que permite a los expertos centrarse en un conjunto reducido de propiedades. Esto se traduciría, por ejemplo, en que para una secuencia de proteína desconocida con este modelo se puede determinar qué tipo de virus es, estudiando su comportamiento en 11 variables en vez de tener que analizar el espectro completo de 256 variables.

5. Conclusiones

La búsqueda de relaciones novedosas y atípicas entre los subtipos del virus de la gripe A, proporciona a los expertos conocimiento novedoso relacionado con este virus que pueda aportar información para ayudarlos en el desarrollo de nuevas terapias o vacunas para este virus.

El conjunto de reglas obtenidas por el algoritmo se podría utilizar por tanto para el desarrollo de nuevas terapias y/o vacunas para mejorar los tratamientos y combatir el virus de la gripe A con un conjunto de solo 11 propiedades.

Acknowledgments.

Este trabajo ha sido subvencionado por el Ministerio de Economía y Competitividad bajo el proyecto TIN201233856, Fondos FEDER, y el por el Plan de Investigación de Andalucía bajo el proyecto TIC-3928, Fondos FEDER.

Referencias

1. M. Atzmueller, F. Puppe, and H. P. Buscher, *Towards Knowledge-Intensive Subgroup Discovery*, Proceedings of the Lernen - Wissensentdeckung - Adaptivität - Fachgruppe Maschinelles Lernen, 2004, pp. 111–117.

2. Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, *The influenza virus resource at the National Center for Biotechnology Information*, Journal of virology **82** (2008), no. 2, 596.
3. C. J. Carmona, P. González, M. J. del Jesus, and F. Herrera, *NMEEF-SD: Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery*, IEEE Transactions on Fuzzy Systems **18** (2010), no. 5, 958–970.
4. C. Chrysostomou, H. Seker, N. Aydin, and P. Haris, *Complex Resonant Recognition Model in Analysing Influenza A Virus Subtype Protein Sequences*, 10th IEEE International Conference on Information Technology and Applications in Biomedicine, 2010.
5. I. Cosic, *Macromolecular bioactivity: is it resonant interaction between macromolecules: Theory and applications*, IEEE transactions on bio-medical engineering **41** (1994), 1101–1114.
6. I. Cosic and E. Pirogova, *Bioactive peptide design using the Resonant Recognition Model*, Nonlinear Biomedical Physics **1** (2007), no. 1, 7.
7. M. J. del Jesus, P. González, F. Herrera, and M. Mesonero, *Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A case study in marketing*, IEEE Transactions on Fuzzy Systems **15** (2007), no. 4, 578–592.
8. M. Fazzolari, R. Alcalá, Y. Nojima, H. Ishibuchi, and F. Herrera, *A review of the application of Multi-Objective Evolutionary Systems: Current status and further directions*, IEEE Transactions on Fuzzy Systems **21** (2013), no. 1, 45–65.
9. F. Herrera, *Genetic fuzzy systems: taxomony, current research trends and prospects*, Evolutionary Intelligence **1** (2008), 27–46.
10. F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus, *An overview on Subgroup Discovery: Foundations and Applications*, Knowledge and Information Systems **29** (2011), no. 3, 495–525.
11. W. Kloesgen, *Explora: A Multipattern and Multistrategy Discovery Assistant*, Advances in Knowledge Discovery and Data Mining, American Association for Artificial Intelligence, 1996, pp. 249–271.
12. N. Lavrac, B. Cestnik, D. Gamberger, and P. A. Flach, *Decision Support Through Subgroup Discovery: Three Case Studies and the Lessons Learned*, Machine Learning **57** (2004), no. 1-2, 115–143.
13. A. Moscona, *Neuraminidase inhibitors for influenza*, New England Journal of Medicine **353** (2005), no. 13, 1363.
14. V. Veljkovic, I. Cosic, B. Dimitrijevic, and D. LalovicC, *Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?*, IEEE Transaction on Biomedical Engineering **32** (1985), no. 5, 337–341.
15. V. Veljkovic, N. Veljkovic, C. P. Muller, S. Mueller, S. Glisic, V. Perovic, and H. Koehler, *Characterization of conserved properties of hemagglutinin of H5N1 and human influenza viruses: possible consequences for therapy and infection control*, BMC Structural Biology **9** (2009).
16. S. Wrobel, *An Algorithm for Multi-relational Discovery of Subgroups*, Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery, LNAI, vol. 1263, Springer, 1997, pp. 78–87.
17. ———, *Inductive logic programming for knowledge discovery in databases*, ch. Relational Data Mining, pp. 74–101, Springer, 2001.
18. L. A. Zadeh, *Soft Computing and Fuzzy Logic*, IEEE Software **11** (1994), no. 6, 48–56.