UNIVERSIDADE FEDERAL DE CIÊNCIAS DA SAÚDE DE PORTO ALEGRE – UFCSPA
CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE

Fernanda Côrte Real Corrêa

Mineração de dados como ferramenta para análise de base de dados de genoma do vírus influenza A

Porto Alegre
2017

Fernanda Côrte Real Corrêa

Mineração de dados como ferramenta para análise de base de dados de genoma do vírus influenza A

Dissertação submetida ao Programa de Pós-Graduação em Ciências da Saúde da Universidade Federal de Ciências da Saúde de Porto Alegre como requisito para a obtenção do título de Mestre em Ciências da Saúde.

Orientador: Dr. Sílvio César Cazella
Coorientadora: Dra. Ana Beatriz Gorini da Veiga

Porto Alegre
2017

Catalogação na Publicação

Corrêa, Fernanda Côrte Real
Mineração de dados como ferramenta para análise de base de dados de genoma do vírus influenza A / Fernanda Côrte Real Corrêa. -- 2017.
103 p. : il., graf., tab. ; 30 cm.

Dissertação (mestrado) -- Universidade Federal de Ciências da Saúde de Porto Alegre, Programa de Pós-Graduação em Ciências da Saúde, 2017.

Orientador(a): Prof. Dr. Sílvio César Cazella ; coorientador(a): Prof. Dr. Ana Beatriz Gorini da Veiga.

1. Vírus da Influenza A. 2. Mineração de Dados. 3. Bases de Dados Factuais. 4. Inteligência Artificial. 5. Biologia Computacional. I. Título.

Sistema de Geração de Ficha Catalográfica da UFCSPA com os dados fornecidos pelo(a) autor(a).

CERTIFICADO

Certificamos que **FERNANDA CÔRTE REAL CORRÊA** apresentou a dissertação de Mestrado no dia 05/09/2017 intitulada "**Mineração de dados como ferramenta para análise de base de dados de genoma do vírus influenza A**" orientada pelo Prof. Silvio César Cazella e coorientada pela: Prof.ª Ana Beatriz Gorini da Veiga junto ao Programa de Pós-Graduação em Ciências da Saúde da Universidade Federal de Ciências da Saúde de Porto Alegre, sendo considerado **aprovada**. Após a homologação da dissertação receberá o título de Mestra em Ciências da Saúde: Epidemiologia e Métodos Diagnósticos.

Porto Alegre, 05 de setembro de 2017.

Gabriela Frank
Secretária Executiva
Pós-Graduação-UFCSPA

Dedico este trabalho aos meus pais,
que sempre me incentivaram a buscar
o conhecimento e sempre me apoiaram
incondicionalmente em todas as
escolhas da minha vida.

AGRADECIMENTOS

Ao prof. Dr. Sílvio César Cazella, meu orientador, pela oportunidade, pela receptividade desde o primeiro dia, por dividir comigo suas experiências e conhecimento, pelo apoio e dedicação constantes. Minha sincera e eterna gratidão.

À prof. Dra. Ana Beatriz Gorini da Veiga, minha coorientadora, pela sabedoria compartilhada e pelos sábios esclarecimentos, meus cordiais agradecimentos.

Aos professores Aline Aver Vanin, Luciano Costa Blomberg e Melissa Santos Fortes por enriquecer os ensinamentos durante as disciplinas do curso.

Aos professores Rafael Andrade Caceres, Pedro Roosevelt Torres Romão, Ygor Arzeno Ferrão, Márcia Rosa da Costa, Marta Quintanilha Gomes, Cleidilene Ramos Magalhães, Caroline Buss e Airton Tetelbom Stein, pela aprendizagem e pela vivência.

Aos servidores Cristiane Mondadori Nolasco e Daniel Rodrigues, pelo excelente atendimento e apoio dados aos alunos desde a seleção até o final do curso.

Às colegas Luiza Silva Vernier, Rafaella Landell de Moura Porto, Carla Thamires Rodriguez Castelli, Gabriela Nunes Maia e Karini Mayer Silva da Cunha pelo companheirismo, pela cumplicidade e pela amizade incondicional.

Aos amigos e colegas de trabalho Janira Prichula, Cristina Almeida da Silva, Andréia Zacharias Mangan, Karen Minatto Eifler, Letícia Eichstaedt Mayer, Milena Meyrer da Silveira e Bruno Barcellos Hervé pelo suporte emocional e técnico, pelo incentivo e pela paciência.

A todos os outros que contribuíram de alguma forma e que não foram aqui citados.

“Muere lentamente quien se transforma en esclavo del hábito, repitiendo todos los días los mismos trayectos, (...) y no le habla a quien no conoce.
(...)
Muere lentamente quien no voltea la mesa cuando está infeliz en el trabajo, quien no arriesga lo cierto por lo incierto para ir detrás de un sueño, quien no se permite por lo menos una vez en la vida, huir de los consejos sensatos.
(...)
Muere lentamente, quien abandona un proyecto antes de iniciarlo, no pregunta de un asunto que desconoce o no respondiendo cuando le indagan sobre algo que sabe.
Evitemos la muerte en suaves cuotas, recordando siempre que estar vivo exige un esfuerzo mucho mayor que el simple hecho de respirar.(...)”

Pablo Neruda

RESUMO

Mineração de dados é o processo de exploração de grandes quantidades de dados com o objetivo de encontrar padrões e correlações capazes de oferecer suporte à tomada de decisões. A mineração de dados é uma ferramenta com grande potencial de aplicabilidade na área de bioinformática, pois permite que volumes robustos de dados sejam processados de forma otimizada. O número de dados biológicos, como por exemplo dados genômicos gerados com as novas tecnologias de sequenciamento, vem crescendo de forma exponencial, sendo necessário cada vez mais o uso de tecnologias computacionais para a interpretação dos mesmos. Os genomas virais constituem uma fonte para o desenvolvimento e uso de novas ferramentas computacionais, devido à vasta quantidade de informação acessível em bases de dados *online*. Este trabalho teve como objetivo analisar uma base pública de dados de genoma de vírus influenza com o emprego de técnicas de mineração de dados. Um banco contendo 232.505 dados de genoma de vírus influenza A e B foi obtido através do site GenBank e pré-processado a fim de eliminar dados incompletos e transformar os dados. Após limpeza, os dados de genoma do vírus influenza A foram minerados com o *software* Weka, com o uso dos algoritmos Apriori e RandomForest para a realização de tarefas de regras de associação e de classificação, respectivamente. A mineração dos dados resultou na correta identificação do vírus influenza A H1N1pdm09. Além disso, os modelos de classificação foram capazes de classificar o subtipo de 74% das amostras de H1N1 (64%) e H3N2 (88%), e de diferenciar o hospedeiro de 77% das amostras aviárias (63%) e humanas (87%). A Mineração de Dados é uma promissora ferramenta para a descoberta de novos conhecimentos na área da saúde, e o *software* Weka possui grande potencial para a aplicação de tarefas de Mineração de Dados nessa área, com capacidade de classificar os dois subtipos mais prevalentes de influenza A e diferenciar entre os dois hospedeiros mais comuns, a partir de dados sequências genômicas do vírus influenza disponíveis em bases de dados públicas.

Palavras-chave: Vírus da Influenza A. Mineração de Dados. Bases de Dados Factuais. Inteligência Artificial. Biologia Computacional.

ABSTRACT

Data mining is a tool with great potential for application in the field of bioinformatics, as it allows extensive volumes of data to be processed in a short period of time. The amount of biological data, such as genomic data generated with the new sequencing technologies, is growing exponentially, and it is increasingly necessary to use computational methodologies for the interpretation of data. Viral genomes are a good source for the development and use of new computational tools due to the vast amount of information available in online databases. This study aimed to analyze a public database of influenza virus genome data using Data Mining techniques. A bank containing 232,505 influenza A and B genome data was obtained from the GenBank website and pre-processed in order to eliminate incomplete data. After cleansing, genome data from influenza A virus were mined using Weka software with Apriori and RandomForest algorithms for association and classification tasks, respectively. Data mining resulted in the identification of influenza A H1N1pdm09. In addition, the classification models were able to correctly classify 74% of the samples of H1N1 (64%) and H3N2 (88%), and also to correctly differentiate the host in 77% of avian (63%) and human (87%) samples. Data Mining presents itself as an excellent tool for knowledge discovery in health sciences and Weka has high potential for application in this field. Weka was able to classify the two most prevalent subtypes of influenza A and also to differentiate between the two most common hosts, starting from a genomic sequences data of influenza virus available in public databases.

Keywords: Influenza A virus. Data Mining. Databases, Factual. Artificial Intelligence. Computational Biology.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BLAST	*Basic Local Alignment Search Tool*
DCBD	Descoberta de Conhecimento em Base de Dados
DNA	ácido desoxirribonucleico
FASTA	*Fast Alignment*
FTP	*File Transfer Protocol*
H1N1pdm09	influenza A H1N1 pandêmico de 2009
HA	hemaglutinina
IAV	vírus influenza A
M1	proteína de matriz 1
M2	proteína de matriz 2
M42	isoforma da M2
NA	neuraminidase
NCBI	*National Center for Biotechnology Information*
NIH	*US National Institutes of Health*
NLM	*National Library of Medicine*
NP	nucleoproteína
NS3	isoforma da NSP1
NSP1	proteína não estrutural 1
NSP2	proteína não estrutural 2
PA	polimerase ácida
PA-N155	isoforma da PA
PA-N182	isoforma da PA
PA-X	isoforma da PA
PB1	polimerase básica 1
PB1-F2	fator de virulência
PB1-N40	isoforma N-terminal truncada da PB1
PB2	polimerase básica 2
RNA	ácido ribonucleico
SNPs	*single nucleotide polymorphisms*
vRNPs	ribonucleoproteínas virais
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Os avanços nas áreas de Biologia Molecular e Bioinformática têm permitido a caracterização de micro-organismos pouco conhecidos, muitos deles patogênicos. Os vírus constituem o grupo de agentes infecciosos com o maior grau de dificuldade de estudo, devido, principalmente, às limitações metodológicas para sua detecção e caracterização, ao fato de muitas infecções virais serem agudas e, também, às altas taxas de mutação e adaptação dos vírus a diferentes hospedeiros. Essas características dos vírus dificultam o diagnóstico e o tratamento das doenças por eles causadas, facilitando a disseminação de epidemias entre humanos. Exemplos são os surtos de Zika vírus e de Febre Amarela que têm ocorrido no Brasil e as epidemias anuais de gripe pelo vírus influenza A[1,2].

Além das epidemias anuais de gripe, o vírus influenza A (IAV) causa, de tempos em tempos, pandemias na população humana. A pandemia de influenza mais conhecida foi a Gripe Espanhola (1918-1919), que levou à morte mais de 50 milhões de pessoas no mundo. Recentemente vivenciamos a pandemia da gripe A de 2009, causada pelo IAV H1N1pdm09, que iniciou com 2 casos em março de 2009 nos Estados Unidos e se espalhou rapidamente pelo mundo, causando a morte de mais de 18 mil pessoas, em mais de 200 países[3].

Novas abordagens baseadas em métodos moleculares e computacionais são fundamentais para avanços no estudo e no controle de doenças infecciosas. Neste contexto, os vírus patogênicos são uma boa fonte de estudo para o desenvolvimento desses novos métodos, haja vista a grande quantidade de informação disponível sobre esses micro-organismos – por exemplo, dados históricos, moleculares e epidemiológicos[4].

As diferentes bases de dados de biologia molecular, as quais contêm uma abundância de informações tais como genomas e proteomas, constituem uma fonte rica em material para pesquisa, acessível de forma fácil, rápida e inteligente. O advento de metodologias que possibilitam a análise desses dados tem facilitado a pesquisa de micro-organismos, sendo possível identificar

mutações, predizer o surgimento de novas cepas com potencial patogênico[5], bem como entender como os agentes patogênicos se espalham geograficamente, entre outras possibilidades[6,7].

Os dados gerados e acumulados nas bases de dados biológicos há muito tornaram-se consistentes e abundantes, gerando um *overload* de conteúdo, de forma que novas tecnologias da informação e técnicas computacionais são necessárias para permitir a análise eficiente e eficaz deste conteúdo. Entre estas novas tecnologias surge o processo de Descoberta de Conhecimento em Base de Dados (DCBD), o qual é composto por etapas, entre elas a etapa denominada de Mineração de Dados[8].

A mineração de dados constitui-se em uma etapa do processo de Descoberta de Conhecimento em Base de Dados que surgiu nos anos 90 com o intuito de correlacionar uma grande quantidade de dados e encontrar padrões em grande velocidade e com extrema facilidade. A quantidade de dados produzida mensalmente por um hospital de grande porte, por exemplo, é muito extensa para ser analisada por métodos estatísticos comuns. A mineração de dados foi desenvolvida, dessa forma, para reunir todos esses dados em uma única análise, de forma que os métodos estatísticos se tornem necessários apenas para alguns dados que demonstrassem correlação nos resultados do procedimento de mineração[9].

Outra vantagem da mineração de dados é que uma determinada instituição pode continuar adicionando dados à base de dados do *software* de mineração que tal instituição utiliza. Assim, a cada nova entrada, o programa irá reprocessar os dados e apresentar novos resultados e padrões instantaneamente, proporcionando uma orientação em tempo real para o médico na hora de proceder com o diagnóstico.

Um dos *softwares* aplicados na Mineração de Dados e amplamente difundido na comunidade científica denomina-se Waikato Environment for Knowledge Analysis (WEKA)[10]. O WEKA apresenta várias aplicações na mineração de dados nas áreas biológicas, como biologia estrutural, caracterização de inibidores de proteínas, análise de classes funcionais de proteínas, identificação gênica, mapeamento fenótipo-genótipo, associação de

SNPs com doenças genéticas, análise de dados de *microarray*, dentre outras aplicações[11].

Até o momento, nenhum trabalho utilizando este *software* para estudos sobre o vírus influenza a partir desta base de dados foi publicado na literatura internacional, apesar da grande quantidade de dados genômicos desse vírus depositadas em bases públicas como o GenBank e o *Influenza Research Database*, Considerando que a Mineração de Dados se apresenta como uma ferramenta capaz de extrair conhecimentos acerca do vírus influenza, a partir de banco de dados proveniente da base GenBank, este trabalho teve como objetivo utilizar o princípio da Mineração de Dados para a análise das informações sobre o vírus influenza A disponíveis publicamente.

Inicialmente, a questão norteadora desta pesquisa era: é possível obter novos conhecimentos através da análise do genoma e do proteoma de vírus patogênicos humanos com o emprego de técnicas de mineração de dados? Como isto não foi possível devido a limitações técnicas da ferramenta de mineração de dados, a questão de pesquisa foi modificada para: é possível obter novos conhecimentos através da análise de banco de dados com informações sobre o genoma do vírus influenza através do uso de técnicas de mineração de dados?

1.1 OBJETIVOS

1.1.1 Objetivo geral

O objetivo geral desta pesquisa é analisar uma base pública de dados de genoma do vírus influenza A com o emprego de técnicas de Mineração de Dados.

1.1.2 Objetivos específicos

- Selecionar uma base de dados pública sobre o vírus influenza para pré-processamento de dados;

- Selecionar tarefas e algoritmos de Mineração de Dados;

- Aplicar algoritmos e tarefas de Mineração de Dados nos dados de genoma selecionados e pré-processados, com o *software* Weka;

- Desenvolver modelos computacionais para a análise de dados, descrição de relacionamentos e classificação de subtipos e hospedeiros do vírus influenza A;

- Realizar o pós-processamento de dados analisando o resultado da Mineração de Dados, em busca de novos conhecimentos acerca do vírus influenza A.

2 REVISÃO BIBLIOGRÁFICA

2.1 VÍRUS INFLUENZA

2.1.1 Classificação, subtipos e nomenclatura

Os vírus influenza são membros da família *Orthomyxoviridae* e são classificados em três gêneros filogeneticamente distintos. O influenza A e o influenza B apresentam oito segmentos de RNA fitas simples senso negativo, enquanto o tipo C possui sete segmentos. Enquanto este último não causa danos significativos à saúde e apresenta pequena variação antigênica, o influenza B – que ocorre apenas em humanos – e o influenza A – que ocorre em humanos, em outros mamíferos e em muitas aves – causam a influenza humana, infecção respiratória aguda considerada uma das mais importantes doenças infecciosas da humanidade, com altas taxas de morbidade e mortalidade[12-14]. A Tabela 1 apresenta a organização genômica dos três gêneros de influenza que afetam humanos.

Tabela 1 – Organização genômica dos vírus influenza.

Segmento	Tamanho*		Peptídeos codificados		
	RNA viral	RNAm	influenza A	influenza B	influenza C
1	2341	2320	PB2	PB1	P1
2	2341	2320	PB1, PB1-F2	PB2	P2
3	2233	2210	PA	PA	P3
4	1778	1756	HA	HA	HEF
5	1565	1539	NP	NP	NP
6	1413	1391	NA	NA, NB	M1, CM2
7	1027	1004, 314, 275	M1, M2	M1, BM2	NS1, NS2
8	890	868, 396	NS1, NS2	NS1, NS2	-

* O tamanho dos RNAs é baseado em sequências do vírus influenza A, podendo variar entre os outros tipos e subtipos, especialmente os segmentos 4, 6 e 8. Fonte: Vedovello[15] adaptado de Wright, Neumann e Kawaoka[16].

Os subtipos de vírus influenza A são classificados pela composição antigênica, baseada nas glicoproteínas de superfície HA e NA. Foram descritos, até o momento, dezoito subtipos de HA (H1 a H18) e onze subtipos de NA (N1 a N11)[17,18]. As Tabelas 2 e 3 mostram os diferentes tipos de hemaglutinina e neuraminidase identificados e as espécies nas quais os mesmos foram detectados.

A nomenclatura do vírus influenza é definida pela Organização Mundial da Saúde, e inclui, na seguinte ordem: o tipo de vírus, baseado na especificidade antigênica do antígeno NP (influenza A, influenza B ou influenza C); o hospedeiro de origem, quando não for humano, ou o local de origem na natureza, para amostras não encontradas em hospedeiros vivos; a origem geográfica do primeiro isolado; o número da linhagem; o ano de isolamento. Para os vírus influenza A, a descrição antigênica é apresentada entre parênteses, incluindo um índice para descrever o subtipo de hemaglutinina (H1 a H18) e um índice para caracterizar cada subtipo de neuraminidase (N1 a N11)[19,20].

Tabela 2 – Subtipos de hemaglutinina e espécies onde foram detectadas.

Subtipo	Seres humanos	Aves domésticas	Porcos	Morcegos	Outros animais
H1	Sim	Sim	Sim		
H2	Sim	Sim	Sim		
H3	Sim	Sim	Sim		Sim
H4		Sim	Sim		Sim
H5	Sim*	Sim	Sim		
H6	Sim*	Sim			
H7	Sim*	Sim			Sim
H8		Sim			
H9	Sim*	Sim	Sim		
H10	Sim*	Sim			
H11		Sim			
H12		Sim			
H13		Sim			
H14		Sim			
H15		Sim			
H16		Sim			
H17				Sim	
H18				Sim	

* Casos raros de transmissão de aves para humanos. Fonte: adaptado de Centers for Disease Control and Prevention[18].

Tabela 3 – Subtipos de neuraminidase e espécies onde foram detectadas.

Subtipo	Seres humanos	Aves domésticas	Porcos	Morcegos	Outros animais
N1	Sim	Sim	Sim		
N2	Sim	Sim	Sim		
N3		Sim			
N4		Sim			
N5		Sim			
N6	Sim*	Sim			
N7	Sim*	Sim			Sim
N8	Sim*	Sim			Sim
N9	Sim*	Sim			
N10				Sim	
N11				Sim	

* Casos raros de transmissão de aves para humanos. Fonte: adaptado de Centers for Disease Control and Prevention[18].

2.1.2 Estrutura do vírus influenza A

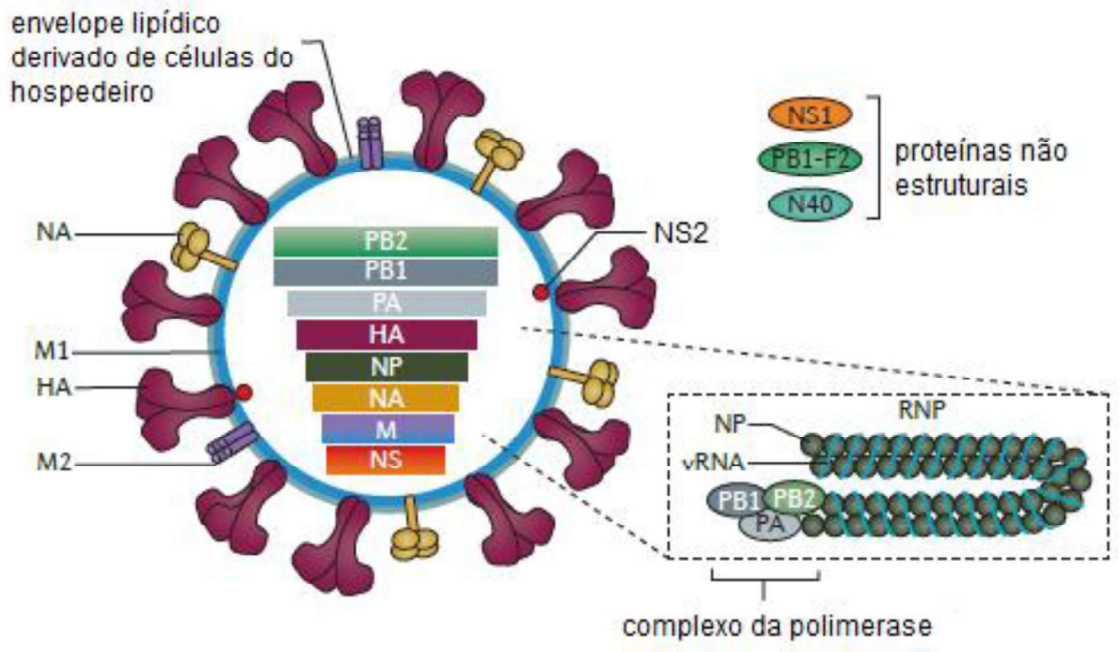
O vírus influenza A (IAV) foi isolado, pela primeira vez, em suínos[21]. Em 1974, foi feita a identificação de uma cepa de origem suína em tecido pulmonar de um indivíduo, demonstrando, pela primeira vez, a capacidade que esses vírus possuem de infectar seres humanos[22].

O IAV pertence à família *Orthomyxoviridae* e possui genoma composto por oito segmentos de RNA de fita simples senso negativo[23]. O genoma codifica onze proteínas virais: proteínas internas (nucleoproteína NP, proteínas de matriz M1 e M2, proteínas não estruturais NSP1 e NSP2, polimerases básicas PB1 e PB2, polimerase ácida PA) e proteínas de superfície (hemaglutinina HA e neuraminidase NA)[24]. Mais recentemente foram descobertas outras proteínas virais que são produzidas por mecanismos moleculares alternativos: PB1-F2, PB1-N40, PA-X, PA-N155, PA-N182, M42 e

NS3[25]. A Figura 1 mostra esquematicamente a estrutura da partícula viral do IAV e suas proteínas estruturais e não estruturais.

O IAV possui alta capacidade de gerar epidemias, especialmente devido a variações antigênicas ocorridas nas proteínas HA e NA[26]. A HA possibilita a ligação do vírus aos resíduos de ácidos siálicos nas glicoproteínas de membrana da célula hospedeira durante a infecção, induzindo a fusão entre as membranas viral e endossomal para a entrada do vírus na célula[27]. A NA facilita a saída das partículas virais do interior das células infectadas. As proteínas do complexo polimerases (PB1, PB2, PA) e a NP são responsáveis pela replicação e transcrição do RNA viral, enquanto a M1 preenche o interior da estrutura viral[28].

Figura 1 – Vírus influenza A: estrutura esquemática da partícula viral e suas proteínas.



Fonte: adaptado de Medina e Garcia-Sastre[29].

2.1.3 Replicação viral

O ciclo de replicação do vírus influenza possui várias fases: ligação do vírus à superfície da célula hospedeira através da hemaglutinina e entrada na célula; introdução das ribonucleoproteínas virais (vRNPs) ao núcleo; transcrição e replicação do genoma viral; exportação das vRNPs a partir do núcleo; montagem do novo vírus na membrana da célula hospedeira; liberação do vírus pela neuraminidase. A clivagem dos ácidos siálicos pela NA permite a disseminação viral em meio extracelular e infecção de novas células[28].

A capacidade de causar epidemias anuais recorrentes e pandemias está relacionada com a alta variabilidade genética e capacidade de adaptação. A fragmentação do genoma permite o rearranjo entre os diferentes segmentos de dois ou mais vírus que infectam uma mesma célula. Além disso, as altas taxas de mutação durante a replicação viral, características de genomas de RNA, contribuem também para o surgimento de novas variantes virais contra as quais a população não está imune[30].

Ainda que todos os tipos de influenza sejam suscetíveis a alterações, o influenza A é o que mais sofre mutações e rearranjos, especialmente nos genes codificantes das proteínas HA e NA. O vírus pode sofrer dois tipos mais comuns de alteração no genoma. As variações antigênicas menores (*antigenic drift*) resultam de um acúmulo de mutações pontuais e envolvem pequenas mudanças na composição das proteínas HA e NA, sendo associadas às epidemias sazonais de influenza[31]. Essas mutações ocorrem devido à infidelidade da RNA polimerase, que possui taxa de uma mutação por genoma a cada replicação, associada à pressão seletiva da imunidade do hospedeiro, resultando em alterações nas proteínas de superfície que são alvo dos anticorpos do hospedeiro[32].

As variações antigênicas maiores (*antigenic shift*) resultam de substituições dos segmentos de RNA de diferentes linhagens de influenza humano e animal, gerando subtipos híbridos com proteínas de superfície misturadas, com potencial de resultar em pandemias. Tais recombinações ocorrem quando uma célula estiver infectada com dois subtipos de influenza

simultaneamente[31]. Evidências sugerem que as pandemias com mudanças nos subtipos de hemaglutinina surgem de rearranjos com vírus influenza A de origem animal[33].

2.1.4 Epidemiologia

As epidemias causadas pelo IAV são caracterizadas por uma rápida dispersão de vírus com novos determinantes antigênicos, que atingem populações suscetíveis por não possuírem imunidade a essas variantes virais, sem apresentar periodicidade ou padrão conhecido de previsibilidade. As epidemias sazonais podem ocorrer durante todo o ano, especialmente em regiões de clima tropical, havendo um aumento da incidência no inverno[34].

As pandemias de influenza decorrem da introdução de um vírus com novo subtipo de HA contra o qual a população humana não possui imunidade. Até o momento, ocorreram quatro grandes pandemias de influenza: a gripe espanhola em 1918, a gripe asiática em 1957, a gripe de Hong Kong em 1968 e a gripe A em 2009[35,36].

A gripe espanhola ocorreu entre 1918 e 1920 e o número reportado de casos foi de aproximadamente 500 milhões de pessoas em todo o mundo. A taxa de mortalidade da gripe espanhola foi superior a 2,5%, causando uma estimativa de 50 a 100 milhões de mortes, níveis elevados se comparados à taxa de 0,1% da gripe sazonal. As pandemias subsequentes e quase todos os casos de epidemias por influenza A, com exceção de infecções provocadas por vírus aviários como H5N1 e H7N7, foram causados por descendentes do vírus de 1918, incluindo variantes do H1N1, H2N2 e H3N2. Esses vírus são compostos por genes oriundos da cepa de 1918, combinados com genes de origem aviária[37,38].

O padrão de morbidade da gripe espanhola foi atípico. Quase metade das mortes relacionadas ao vírus influenza na pandemia de 1918 ocorreu em jovens adultos saudáveis de 20 a 40 anos de idade. A doença se apresentava com progressão rápida resultando em falência múltipla de órgãos, e as taxas

de mortalidade da gripe e pneumonia entre 15 e 34 anos de idade foram 20 vezes maiores do que em anos anteriores[39].

Após o primeiro isolamento do vírus influenza A em seres humanos em 1933[40], a vigilância para a ocorrência de novas pandemias aumentou. Entretanto, após a pandemia de gripe de 1918, o vírus voltou a apresentar o padrão habitual regional e de menor virulência nas décadas de 1930, 1940 e início dos anos 50[33].

Em 1957, uma nova pandemia de gripe iniciou na China. O vírus foi rapidamente identificado como influenza A. Entretanto, testes revelaram a presença de proteínas hemaglutinina e neuraminidase diferentes das encontradas anteriormente em humanos[41]. A doença foi registrada em Hong Kong, Singapura, Japão, Indonésia, Filipinas, entre outros. A gripe asiática, causada pelo vírus A/Singapura/1/57(H2N2), foi responsável por, aproximadamente, um milhão de óbitos em todo o mundo. Através de viagens de navio, o vírus se disseminou para Estados Unidos, Holanda e Austrália[33,42].

Em 1968, uma nova pandemia surgiu no sudeste da Ásia, causada pelo subtipo H3N2. O primeiro surto teve início em Hong Kong e rapidamente avançou por toda a Ásia. Entre 1968 e 1969, a pandemia provocada pela cepa A/Hong Kong/1/68(H3N2) resultou em cerca de um milhão de óbitos em todo o mundo e acometeu 15% da população de Hong Kong (cerca de 500 mil pessoas). Nos Estados Unidos, ocorreram aproximadamente 34 mil óbitos[43].

Em março e abril de 2009, uma nova cepa de influenza A H1N1 surgiu no México e nos Estados Unidos. Esse vírus continha seis fragmentos de RNA derivados de cepas recombinantes de H3N2 e/ou H1N2 de origem norte americana, contendo genes de origem humana, aviária e suína. Por outro lado, os demais genes (neuraminidase e matriz) eram origem suína provenientes da Eurásia[44-46].

O vírus influenza responsável pela pandemia de 2009 foi inicialmente classificado como A/California/04/2009, e foi, posteriormente, chamado de influenza A H1N1pdm09. O vírus nunca havia circulado entre humanos. Mesmo com a implantação de medidas de contenção, o vírus se disseminou

rapidamente por todo o mundo e, em 11 de junho de 2009, a Organização Mundial de Saúde declarou oficialmente a pandemia de gripe[47,48].

O H1N1pdm09 substituiu a cepa H1N1 sazonal que circulava previamente, e circula atualmente com o influenza H3N2 sazonal. Esses vírus fazem parte da composição das vacinas anuais contra a gripe[49].

2.2 BIOLOGIA COMPUTACIONAL

2.2.1 Análise de dados de genomas e proteomas

Até o final dos anos 1960, já existiam variadas técnicas computacionais para análise de estrutura, função e evolução moleculares, bem como bancos de dados de sequências proteicas. Mesmo sem os benefícios de supercomputadores ou redes de computadores, os cientistas da época conceberam importantes conceitos e fundamentos técnicos que servem de base para a bioinformática até hoje[50].

Nas décadas seguintes, foram criados os primeiros algoritmos para alinhamento de sequências e as bases de dados de acesso público. Além disso, foram aprimorados os sistemas de busca em bancos de dados e desenvolvidas ferramentas para anotação e comparação de genomas[51].

Desde o final da década de 1980, o termo “bioinformática” tem sido utilizado principalmente em referência a métodos computacionais para análise comparativa de dados de genomas. No entanto, o termo foi originalmente criado para definir de forma mais ampla o estudo de processos informáticos em sistemas bióticos[52].

Atualmente, existem numerosos algoritmos e *softwares* para uso em bioinformática. Entre as suas funções, é possível destacar: comparação de biossequências a fim de encontrar trechos semelhantes entre elas; montagem de fragmentos de DNA de forma a reconstituir o trecho de DNA do qual os mesmos originam-se; mapeamento físico de cromossomos ou DNA; construção de árvores filogenéticas para esclarecer o histórico evolutivo dos organismos;

predição de estruturas tridimensionais[53]. A otimização de bancos de dados, o aprimoramento de algoritmos rápidos de classificação e agrupamento e os *softwares* de mineração de dados são as principais áreas de desenvolvimento atuais com aplicações na bioinformática[54].

2.3 MINERAÇÃO DE DADOS

2.3.1 Descoberta de Conhecimento em Bases de Dados

A Descoberta de Conhecimento em Bases de Dados (DCBD) constitui-se no processo de explorar grandes quantidades de dados à procura de padrões escondidos e consistentes, que podem ser descobertos através de tarefas de mineração de dados tais como: regras de associação, regras de classificação, clusterização ou análise de séries temporais. Desta forma, torna-se possível detectar relacionamentos sistemáticos entre variáveis, apresentando conhecimento novo a partir de subconjuntos de dados[8].

Em geral, o processo de DCBD consiste em uma iteração das etapas abaixo. A Figura 2 esquematiza a DCBD de forma simplificada.

1. Seleção dos dados: a primeira etapa consiste em escolher qual conjunto de dados será submetido ao processo. Desta forma, é selecionado um conjunto de dados alvo ou um subconjunto de variáveis ou amostras de dados.

2. Pré-Processamento: é a etapa onde os dados são preparados para serem apresentados às técnicas de mineração de dados. Os dados são selecionados (de acordo com sua relevância), purificados (são removidas as inconsistências e incompletude dos dados) e pré-processados (formatados de uma maneira adequada para a mineração de dados). Este passo é realizado sob a supervisão e conhecimento de um especialista na área, pois o mesmo é capaz de definir quais dados são importantes, assim como o que fazer com os dados antes de utilizá-los no data mining.

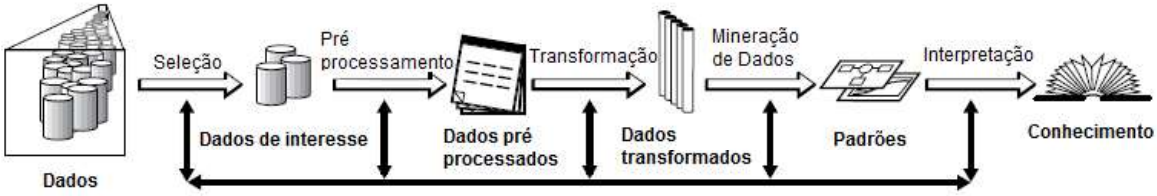
3. Transformação: nesta fase, os dados são convertidos em um formato adequado para serem processados pelos algoritmos de mineração. É neste

momento que pode ser feita uma redução no número de variáveis, sumarizando dos dados que serão submetidos à mineração.

4. Mineração de dados: é onde os dados preparados são processados, ou seja, é onde se faz a mineração dos dados propriamente dita. Nesta fase, o algoritmo escolhido é aplicado sobre os dados a fim de se descobrir padrões interessantes.

5. Pós-Processamento: constitui-se a etapa de visualização e interpretação dos dados, onde o resultado da mineração é avaliado, visando determinar se algum conhecimento adicional foi descoberto, assim como definir a importância dos fatos gerados, sob a supervisão de um especialista na área.

Figura 2 – Descoberta de conhecimento em base de dados.



Fonte: adaptado de Fayyad, Piatetsky-Shapiro e Smyth[9].

Atualmente, a mineração de dados é uma das tecnologias mais utilizadas para extração de conhecimento a partir de bancos de dados, tanto no meio comercial quanto no meio científico. A mineração de dados permite a exploração e análise, de forma automática ou semiautomática, de grandes quantidades de dados de forma a identificar padrões e regras significativos[55].

Podem haver três níveis de mineração de dados de genoma. O mais simples é uma análise em profundidade do resultado a partir de uma única informação, que pode começar com um gene ou marcador, ou mapeando uma sequência do genoma. O próximo nível de mineração envolve a seleção de um conjunto de genes ou *loci* que atendem a um critério ou combinação de critérios, seguida pela captura de dados para análise em profundidade. Para um estudo mais detalhado, o processamento em lotes e integração com pacotes estatísticos é mais adequado[56].

2.3.2 Tarefas de Mineração de Dados

A mineração de dados pode ser categorizada como descritiva (aprendizado não-supervisionado) ou preditiva (aprendizado supervisionado). A mineração descritiva tem como objetivo pesquisar conjuntos volumosos de dados em busca de relacionamentos, padrões, tendências, grupamentos ou dados atípicos (*outliers*). Por outro lado, a mineração preditiva se baseia na construção de modelos de regressão, classificação, reconhecimento de padrões ou tarefas de aprendizado de máquina, e é capaz de avaliar a precisão preditiva desses modelos e procedimentos quando aplicada a novos dados[57].

As tarefas de mineração de dados consistem nas especificações daquilo que estamos querendo buscar nos dados, que tipo de regularidades ou categorias de padrões de interesse, ou ainda que tipo de padrão é relevante ou não. A análise de regras de associação e a análise de classificação e predição são exemplos de tarefas de mineração de dados[58].

As tarefas de associação são técnicas de mineração utilizadas para descobrir as relações entre um grande conjunto de variáveis e um conjunto de dados, identificando quais atributos estão relacionados. O objetivo da mineração, nesse caso, é gerar regras que identifiquem a presença de um conjunto de dados implicando na presença de outro conjunto de dados, ou seja, encontrar relacionamentos ou padrões frequentes entre os dados[59].

As tarefas de classificação propõem-se a identificar a qual classe um determinado registro pertence. Nesta tarefa, um conjunto de registros é analisado, com cada registro já contendo a indicação à qual classe pertence. Desta forma, a máquina ‘aprende’ como classificar um novo registro, e passa a ser capaz de predizer em qual categoria um novo dado se encaixa[58].

2.3.3 Algoritmos e métricas

O algoritmo Apriori é um dos algoritmos mais conhecidos para mineração por regras de associação[60,61]. O algoritmo emprega busca em profundidade e

gera conjuntos de itens candidatos (padrões) de *k* elementos a partir de conjuntos de itens de *k-1* elementos. Os padrões não frequentes são eliminados. Toda a base de dados é rastreada e os conjuntos de itens frequentes são obtidos a partir dos conjuntos de itens candidatos. O objetivo deste algoritmo é procurar relações entre os dados enquanto eles são separados. Simultaneamente, o algoritmo calcula valores correspondentes a suporte (*support*) e confiança (*confidence*), que refletem respectivamente a utilidade e a confiabilidade da regra descoberta[59].

Os parâmetros confiança e suporte são essenciais para o funcionamento do algoritmo. Eles determinam diretamente tanto a quantidade como a qualidade das regras geradas. O suporte de um item é calculado pelo percentual de vezes em que ele ocorre em relação ao conjunto total de dados. A confiança representa o percentual de ocorrência da regra gerada, avaliando o quão forte é uma regra. Ela indica o quanto a ocorrência do antecedente de uma regra pode influenciar na ocorrência do consequente da regra[62].

Para calcular o nível de dependência entre itens, pode-se utilizar mais medidas tais como *lift*, *leverage* e convicção (*conviction*). A medida de avaliação *lift* permite eliminar regras com confiança elevada, mas com pouco interesse. O *lift* avalia se dois itens são positivamente ou negativamente independentes, e também determina quando dois itens são independentes entre si[58].

A medida *leverage* define a diferença entre a proporção de exemplos cobertos, simultaneamente, pelo antecedente e pelo consequente da regra e a proporção de exemplos que seriam cobertos se o antecedente e o consequente fossem independentes. Por fim, a medida convicção (*conviction*) permite medir a independência do item antecedente face ao consequente da regra[59].

A partir de uma categorização pré-determinada de registros de uma base de dados, é possível que sejam realizadas predições de comportamentos futuros. Os métodos utilizados para realizar estas tarefas de classificação incluem árvores de confusão, redes bayesianas, funções e classificação por meio de regras[63].

O algoritmo ZeroR efetua a classificação por meio de regras, predizendo qual valor nominal é mais frequente na base de dados de treinamento. Os resultados são apresentados através de uma matriz de confusão contendo o percentual de acerto para um determinado atributo[63].

As redes bayesianas são representações elaboradas a partir de formalizações matemáticas. O algoritmo NaiveBayes calcula a probabilidade que uma amostra desconhecida tem de pertencer a cada uma das classes possíveis, ou seja, prediz a classe mais provável da amostra[64].

O algoritmo SimpleLogistic classifica os dados a partir de funções, criando modelos de regressão logística linear. A regressão logística é uma abordagem para a predição de um desfecho dicotômico, em que analisa a relação entre uma ou mais variáveis que podem predizer a probabilidade de ocorrência de um determinado dado ou desfecho[65,66].

Os algoritmos Random Forest, RandomTree e J48 são utilizados para a realização de tarefas de classificação, fazendo uso do método de árvores de decisão. Esses algoritmos geram estruturas em formato de árvore, cujas ramificações representam as decisões a partir das quais são geradas regras de classificação dos dados[10].

Os algoritmos J48 e RandomTree realizam tarefas de classificação através da geração de uma única árvore de decisão. O algoritmo J48 possui o objetivo de construir uma árvore a partir de um conjunto de dados, onde o atributo mais significativo é considerado a raiz da árvore. O algoritmo RandomTree considera apenas alguns atributos escolhidos aleatoriamente para cada nó da árvore gerada[8].

O algoritmo RandomForest é um dos algoritmos mais utilizados para mineração através de tarefas de classificação. Enquanto os usuais fazem uma construção total de uma estrutura a partir de uma base de dados, o RandomForest tem como objetivo criar várias árvores de decisão usando um subconjunto de atributos selecionados aleatoriamente a partir do conjunto original, contendo todos os atributos[67].

Com a quebra dos dados e construção de vários subconjuntos, uma árvore de decisão é construída. Cada árvore é construída utilizando uma

amostra aleatória inicial dos dados, e a cada divisão desses dados, um subconjunto aleatório de atributos é utilizado para a escolha dos atributos mais informativos. Com este procedimento, os atributos são aplicados aos nós de cada uma das árvores criadas, formando uma 'floresta aleatória' ou *random forest*[68].

Após a criação dos conjuntos de árvores, é possível prosseguir com a classificação dos dados. O algoritmo escolhe um subconjunto de árvores que possui melhor lógica. Para cada subconjunto é dado um voto sobre qual classe o atributo-chave deve pertencer, e este voto possui um 'peso' afetado pela igualdade entre as árvores. Quanto menor for a similaridade entre duas árvores, mais força cada uma delas terá individualmente, e mais preciso será o seu 'peso' na composição da árvore final[69].

O algoritmo RandomForest apresenta várias vantagens sobre outros algoritmos de classificação por árvore de decisão (*trees*). Além de possuir uma técnica exata, ele possui alto nível de acurácia e boa taxa de acertos quando testado em diferentes conjuntos de dados, sendo mais poderoso quando comparado a outros algoritmos de classificação. O algoritmo permite uma classificação aleatória das árvores sem intervenção humana, reduzindo erros, é menos sensível a ruídos e evita sobreajuste de dados (*overfitting*). Além disso, é capaz de lidar com grandes *datasets* e com grandes números de atributos simultaneamente[70].

2.3.4 *Software* Weka

O principal *software* aplicado à Mineração de Dados, amplamente difundido na comunidade científica, denomina-se Waikato Environment for Knowledge Analysis (WEKA). Weka é um *software* livre do tipo *open source* para Mineração de Dados, desenvolvido em Java, dentro das especificações da General Public License (GLP). O *software* foi desenvolvido por um grupo de pesquisadores da Universidade de Waikato, Nova Zelândia[10]. Ao longo dos anos se consolidou como a ferramenta de Mineração de Dados mais utilizada em ambiente acadêmico. Dentro do processo de Descoberta de Conhecimento

em Banco de Dados (DCBD), a ferramenta permite trabalhar o pré-processamento, a mineração de dados propriamente dita e o pós-processamento. O *software* também permite a modelagem do fluxo do processo de descoberta através da ferramenta de *Knowledge Flow*[71].

2.4 BANCO DE DADOS DE GENOMA

2.4.1 GenBank

Os bancos de dados biológicos representam, hoje, uma das principais ferramentas de suporte para pesquisadores de diversas áreas biológicas e biomédicas, incluindo Biologia Molecular, Genética, Microbiologia, Imunologia, Bioinformática, dentre outras. Nestes bancos são feitos cadastros de sequências, anotações biológicas e inclusão de dados relacionados, além de consultas visando o levantamento de dados para análises[53,72].

Os bancos podem ser classificados de acordo com as informações biológicas que armazenam. O conteúdo disponível inclui, principalmente: sequências (de nucleotídeos ou de proteínas) e anotações sobre as mesmas; proteínas e informações sobre as respectivas funções; estruturas de moléculas de proteínas (secundárias e terciárias); taxonomia; bibliografia na área de biologia molecular[73].

Os bancos de sequências de nucleotídeos reúnem as sequências propriamente ditas, além de anotações contendo dados de características biológicas relevantes sobre elas, tais como organismo às quais pertencem, sequências codificadoras de proteínas, função, fenótipo. No caso de patógenos, são incluídas informações como nome da cepa e outras especificações, hospedeiro e características (sexo, idade) do mesmo, localização geográfica, data, entre outros[74].

Um dos bancos de dados biológicos mais completos disponíveis atualmente é o GenBank, criado, distribuído e mantido pelo *National Center for Biotechnology Information* (NCBI), uma divisão da *National Library of Medicine*

(NLM), localizada no campus central do *US National Institutes of Health* (NIH) em Bethesda, Maryland, EUA. Este banco implementa o arquivamento dos objetos utilizando dados semiestruturados. Além disso, os dados complexos também podem ser armazenados à parte em formatos específicos a fim de permitirem manipulação por algoritmos especiais, tais como FASTA e BLAST[75].

O GenBank é o mais importante repositório amplo de sequências de nucleotídeos. O histórico do volume de sequências armazenadas na base do GenBank demonstra que, a cada ano, o número sequências e bases armazenadas cresce cerca de 70% por ano. O sistema permite que a quantidade de informações, bem como a inclusão ou alteração de atributos, seja alterada frequentemente[53].

O GenBank atribui registros sequenciais a divisões de dados, baseando-se na taxonomia de origem ou estratégia de sequenciamento usada para obter os mesmos. Atualmente, há 12 divisões taxonômicas e cinco divisões de alto rendimento. A divisão VRL, de vírus, possui uma taxa de crescimento anual de 19,2%[76].

O NCBI distribui as versões dos bancos de dados do GenBank para *download* por meio do servidor de arquivos FTP no *site* ftp://ftp.ncbi.nlm.gov/genbank. A versão completa no formato de arquivo simples está disponível como um conjunto de arquivos compactados além de um conjunto não cumulativo de atualizações[77].

2.5 A INTELIGÊNCIA ARTIFICIAL NO ESTUDO DO VÍRUS INFLUENZA

2.5.1 Aplicação de algoritmos na análise de dados epidemiológicos

O Aprendizado de Máquina é uma subárea da Inteligência Artificial que visa o desenvolvimento de técnicas computacionais capazes de adquirir conhecimento de forma automática. Um sistema de aprendizado é um algoritmo que toma decisões baseado em experiências acumuladas por meio da solução bem-sucedida de problemas anteriores[78].

A inteligência artificial articula estratégias para simular o comportamento da natureza, entre outros, como método de solução de problemas computacionais. Neste sentido, a mineração de dados busca descobrir generalizações nos dados, e como uma instância de aprendizado, a inteligência artificial contribui na pesquisa e modelagem dessas estratégias de simulação[79].

O crescimento exponencial do volume das bases de dados sobre agentes patogênicos levou a uma melhor compreensão da capacidade de mutação e de atingir diferentes hospedeiros desses agentes. Porém, ainda há muitas perguntas a ser respondidas.

A aplicação de algoritmos na análise de dados epidemiológicos tem sido realizada com sucesso por pesquisadores. Estudos já foram realizados para predizer novas epidemias de dengue e malária, por exemplo. Entretanto, mesmo com trabalhos publicados apresentando resultados satisfatórios, a quantidade de artigos publicados ainda é muito pequena[80,81].

2.5.2 Aplicação de algoritmos na predição de epidemias de gripe

Um algoritmo integrado de associação e classificação foi utilizado para descobrir os fatores associados à transformação de sequências de vírus influenza A não-pandêmicos em sequências de IAV pandêmicas, e posteriormente desenvolver um sistema eficiente para predição de epidemias

de gripe. Neste estudo, foi utilizado um *dataset* contendo 5.373 sequências de HA de cepas de H1N1 pandêmicas e não-pandêmicas isoladas em 2009. Os pesquisadores desenvolveram um *software* para a predição de pandemias de influenza, baseado na descoberta de regras de associação entre pontos de mutação durante a evolução da gripe pandêmica[82].

Em outro estudo, os algoritmos ARIMA e RandomForest foram aplicados em modelos de séries temporais para analisar retrospectivamente dados de incidência de surtos de IAV H5N1 aviária no Egito. O algoritmo RandomForest apresentou melhores resultados na capacidade preditiva, e os pesquisadores concluíram que o algoritmo é efetivo para predizer epidemias de H5N1 no Egito[83].

2.5.3 Aplicação de algoritmos na predição de tropismo de hospedeiro

Em um estudo, foram construídos modelos computacionais para 11 proteínas do IAV utilizando o algoritmo RandomForest para a predição de tropismo do hospedeiro. Os modelos de predição foram criados e treinados com 67.940 sequências de proteínas isoladas a partir de amostras tanto de aves como de seres humanos, obtidas da base *Influenza Research Database* e transformadas em vetores criados a partir das propriedades físico-químicas dos aminoácidos. Os resultados foram modelos de predição altamente precisos, capazes de determinar o tropismo de proteínas de IAV individuais por um hospedeiro específico[84].

2.5.4 Aplicação de algoritmos na predição de resposta a vacinas

Alguns trabalhos já foram realizados para desenvolver modelos computacionais capazes de predizer os resultados de campanhas de vacinação contra a gripe. Em um estudo preliminar com o uso de algoritmos de rede neural, foi construída uma base de dados médicos de 90 pacientes para criar um modelo de vacinação que pudesse ser aplicado na prática da atenção

primária. A base continha dados como titulação prévia de anticorpos contra cepas de influenza A e B, número de vacinações prévias, idade, resposta à vacinação, entre outros, totalizando 27 variáveis[85].

Em um estudo similar, foram selecionados 93 pacientes vacinados contra a gripe. Um banco de dados contendo 52 parâmetros foi gerado e utilizado para construir um modelo de predição de resultados de vacinação contra o vírus influenza através de algoritmos de regressão logística. Os pesquisadores mostraram que é possível desenvolver modelos úteis para predição de resposta à vacinação através da seleção adequada de atributos[86].

2.5.5 Aplicação de algoritmos na pesquisa de antivirais

Em outro estudo, foram construídos modelos de classificação para avaliar possíveis inibidores da neuraminidase, com os algoritmos Support Vector Machine e Naïve Bayesian. Os modelos foram criados utilizando compostos sabidamente ativos e inativos para predizer a atividade inibitória de 15.600 compostos. Após análise computacional, os melhores compostos foram testados *in vitro* para verificar a atividade contra H1N1 e H3N2, resultando na descoberta de 9 novos inibidores de neuraminidase[87].

3 METODOLOGIA

3.1 Pré-processamento

Esta pesquisa caracteriza-se como uma pesquisa exploratória segundo seu objetivo, sendo de natureza aplicada e de abordagem quantitativa. O método de pesquisa envolve experimentação utilizando o processo de Descoberta de Conhecimento em Base de Dados (DCBD) e consequentemente a técnica de Mineração de Dados.

O banco de dados de genomas do vírus influenza genomeset.dat foi obtido por meio do site GenBank[88]. Este *dataset* apresentava dados de cepas dos vírus influenza A e B identificadas, sequenciadas e registradas até o dia 05 de fevereiro de 2016. O total de registros do banco era de 232.505 dados, incluindo cepas isoladas oriundas de um período entre 1902 e 2016.

Na Tabela 4, está uma amostra dos registros do banco de dados genomeset.dat, composto por 11 campos: código de acesso no GenBank; hospedeiro; número do segmento do genoma viral; subtipo; local; data; comprimento/tamanho da sequência; nome; idade do hospedeiro; sexo do hospedeiro; número de registro da cepa. Neste exemplo está incluído, também, o vírus influenza B.

Tabela 4 – Amostra dos registros do banco de dados genomeset.dat.

código	hospe-deiro	nº	sub-tipo	Local	data	tama-nho	Nome	idade	sexo	nº da cepa
M14880	Human	1	-	USA	1940	2368	Influenza B virus (B/Lee/1940)	-	-	14656
AF101982	Human	2	-	USA	1940	2313	Influenza B virus (B/Lee/1940)	-	-	14656
AF102017	Human	3	-	USA	1940	2204	Influenza B virus (B/Lee/1940)	-	-	14656
K00423	Human	4	-	USA	1940	1882	Influenza B virus (B/Lee/1940)	-	-	14656
K01395	Human	5	-	USA	1940	1841	Influenza B virus	-	-	14656
J02095	Human	6	-	USA	1940	1557	Influenza B virus	-	-	14656
J02094	Human	7	-	USA	1940	1191	Influenza B virus	-	-	14656
J02096	Human	8	-	USA	1940	1096	Influenza B virus	-	-	14656
CY040652	Human	1	H1N1	USA	30/04/09	2292	Influenza A virus (A/New York/3194/2009(H1N1))	9Y	-	10035 50
CY026154	Human	1	H3N2	USA	08/01/20 07	2301	Influenza A virus (A/Colorado/UR06-0023/2007(H3N2))	3Y	F	10025 97
CY033364	Avian	4	H3N5	USA	05/12/20 07	1765	Influenza A virus (A/northern shoveler/California/HKW F1199/2007(H3N5))	Hatch Year	M	10028 11
GQ200292	Human	1	H1N1	China	10/05/09	2328	Influenza A virus (A/Shandong/1/2009(H1N1))	-	-	10035 98
GU477553	Avian	2	H5N1	China	2009/06	2274	Influenza A virus (A/great black-headed gull/Qinghai/8/2009(H5N1))	-	-	10214 21
AJ404626	Human	4	H9N2	Hong Kong	1999	1714	Influenza A virus (A/Hong Kong/1073/99(H9N2))	-	-	14892
AB450626	Avian	7	H5N1	Thailand	2004	982	Influenza A virus (A/chicken/Kalasin/NIAH 316/2004(H5N1))	-	-	10028 17
CY014686	Avian	1	H11N6	UK	1956	2341	Influenza A virus (A/duck/England/1/1956(H11N6))	-	-	24461
GU186781	Avian	8	H7N7	Italy	1902	865	Influenza A virus (A/chicken/Brescia/1902(H7N7))	-	-	10048 29
EU053149	Swine	5	H1N2	Germany	2000	1534	Influenza A virus (A/swine/Bakum/1832/2000(H1N2))	-	-	10333 97
CY103880	Bat	8	H17N10	Guatemala	2009/05	895	Influenza A virus (A/little yellow-shouldered bat/Guatemala/153/2009(H17N10))	-	-	10126 34

Foi realizada a limpeza do arquivo, através da remoção de dados incompletos e duplicados, e remoção de atributos não interessantes para o processo (códigos numéricos, idade, sexo), seguida por redução do atributo data (informações referentes a dia/mês foram removidas, sendo mantido apenas o ano). Dados sobre vírus influenza de tipos B e C foram removidos. Em seguida, foram selecionados os dados relativos ao período 2005-2015, visto que esse período contém dados mais completos e uniformes.

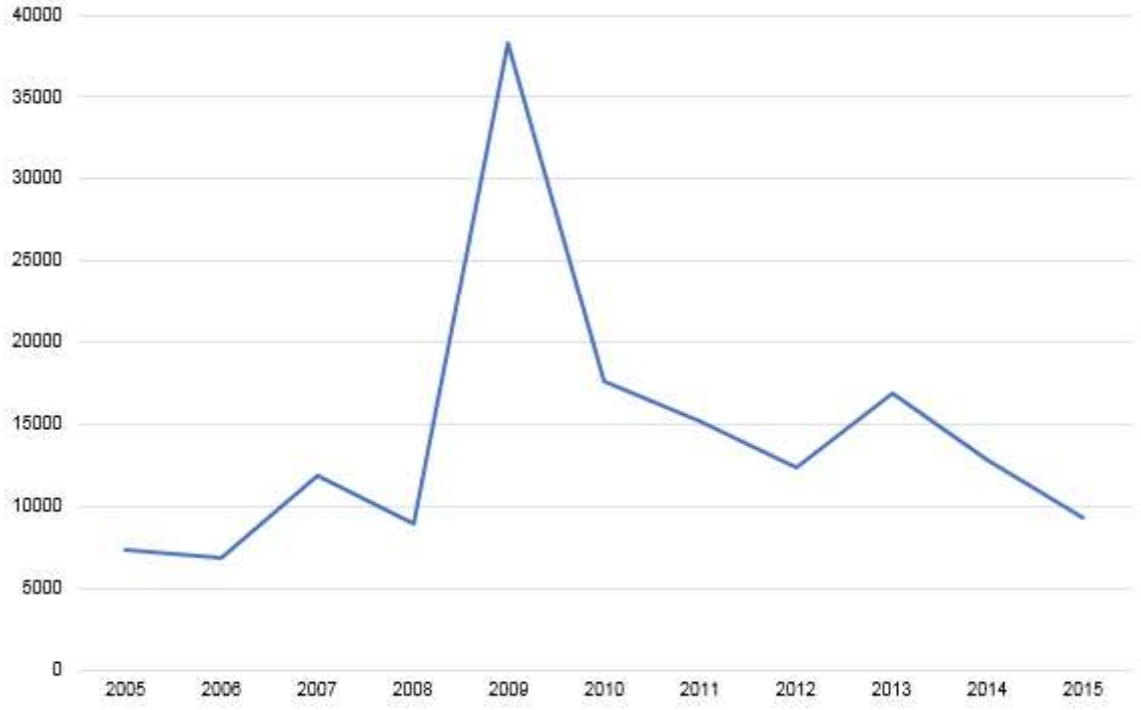
O arquivo gerado após a fase de pré-processamento possui 157.639 instâncias. A Tabela 5, contém uma amostra dos registros do banco de dados genomeset.dat após processamento, composto por 6 campos: hospedeiro (fonte); número do segmento do genoma viral (fragmento); subtipo; local; data (ano); comprimento/tamanho da sequência (tamanho).

Os Gráficos 1 a 6 demonstram a distribuição dos dados de genoma do vírus influenza A após a realização da etapa de pré-processamento, entre os anos de 2005 a 2015. O Gráfico 1 consiste na distribuição de IAV por ano; o Gráfico 2 demonstra a distribuição de dados por fragmento de genoma de IAV; o Gráfico 3 evidencia a distribuição do subtipo H1N1; o Gráfico 4 reflete a distribuição de IAV por hospedeiro; o Gráfico 5 caracteriza a distribuição de dados por país; o Gráfico 6 reproduz a distribuição de dados de IAV por subtipo de vírus. Versões expandidas e mais completas dos gráficos estão disponíveis como apêndices deste trabalho.

Tabela 5 – Amostra dos registros do banco de dados após processamento.

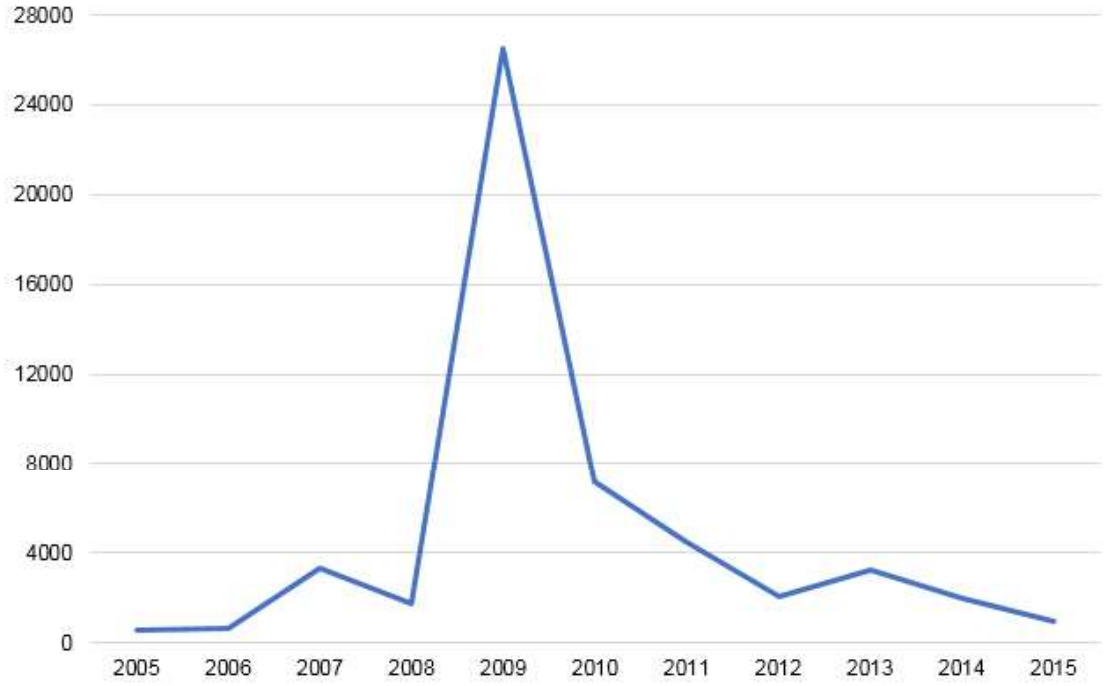
Fonte	Fragmento	Subtipo	Local	Ano	Tamanho
Avian	8	H11N1	USA	2005	841
Avian	7	H11N1	USA	2005	990
Avian	6	H11N1	USA	2005	1420
Avian	5	H11N1	USA	2005	1530
Avian	4	H11N1	USA	2005	1727
Avian	3	H11N1	USA	2005	2161
Avian	2	H11N1	USA	2005	2284
Avian	1	H11N1	USA	2005	2307
Avian	8	H11N2	Netherlands	2005	855
Avian	7	H11N2	Netherlands	2005	985
Avian	6	H11N2	Netherlands	2005	1427
Avian	5	H11N2	Netherlands	2005	1530
Avian	4	H11N2	Netherlands	2005	1705
Avian	3	H11N2	Netherlands	2005	2177
Avian	2	H11N2	Netherlands	2005	2284
Avian	1	H11N2	Netherlands	2005	2287
Avian	8	H11N2	USA	2005	855
Avian	7	H11N2	USA	2005	984
Avian	6	H11N2	USA	2005	1434
Avian	5	H11N2	USA	2005	1530
Avian	4	H11N2	USA	2005	1727
Avian	3	H11N2	USA	2005	2182
Avian	2	H11N2	USA	2005	2291
Avian	1	H11N2	USA	2005	2295
Human	8	H1N1	Guam	2009	850
Human	7	H1N1	Guam	2009	987
Human	6	H1N1	Guam	2009	1420
Human	5	H1N1	Guam	2009	1525
Human	4	H1N1	Guam	2009	1734
Human	3	H1N1	Guam	2009	2175
Human	1	H1N1	Guam	2009	2293
Human	2	H1N1	Guam	2009	2303

Gráfico 1 – Distribuição de dados de influenza A por ano entre 2005 e 2015.



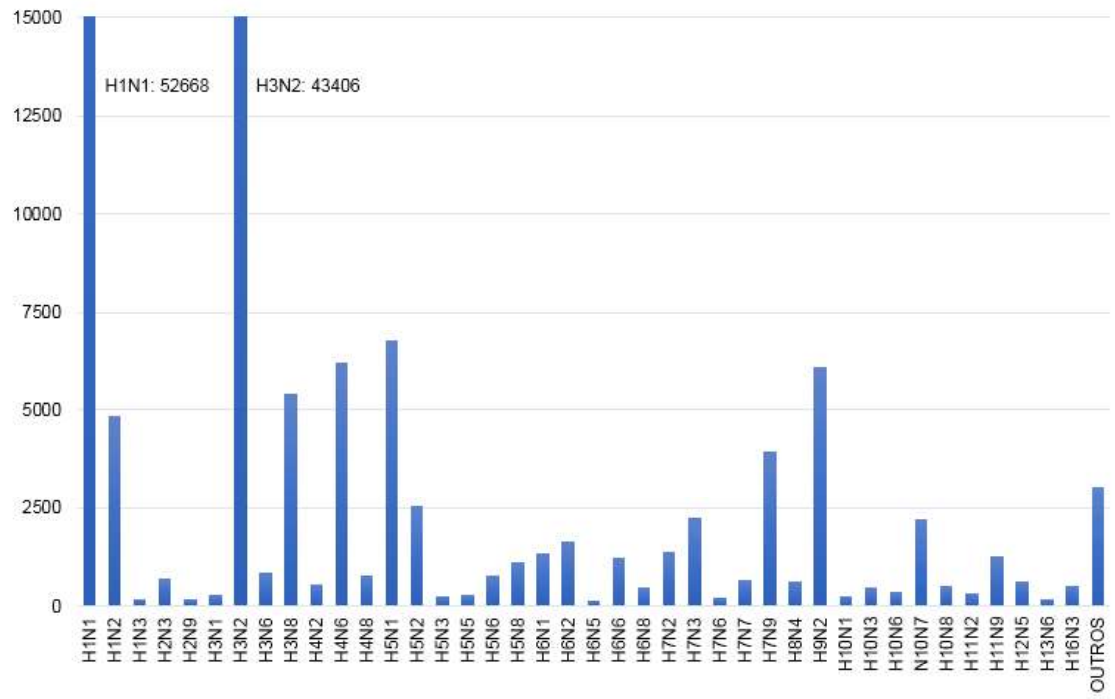
Fonte: produção do próprio autor.

Gráfico 2 – Distribuição de dados de influenza A H1N1 entre 2005 e 2015.



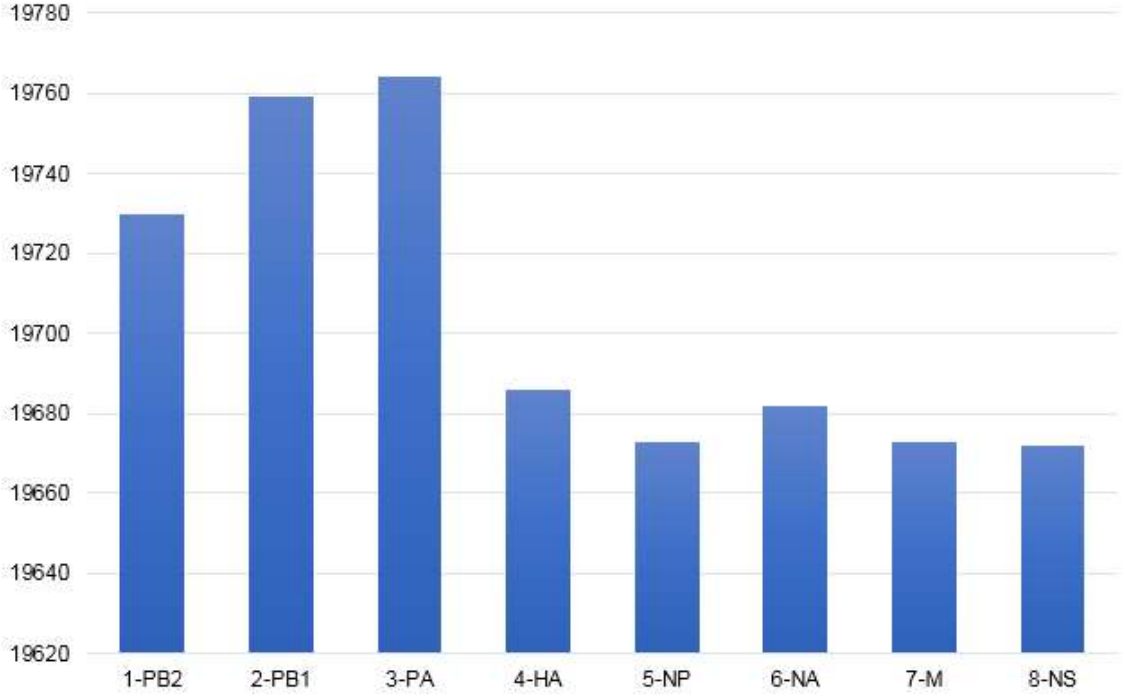
Fonte: produção do próprio autor.

Gráfico 3 – Distribuição de dados de vírus influenza A por subtipo entre 2005 e 2015.



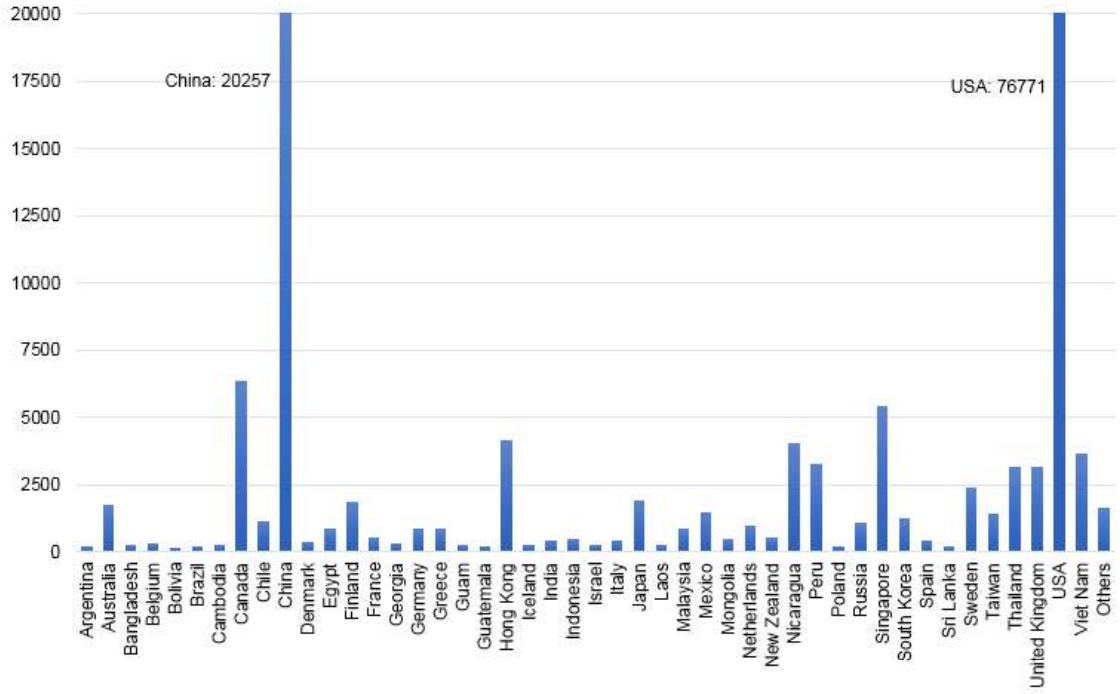
Fonte: produção do próprio autor.

Gráfico 4 – Distribuição de dados por fragmento de genoma entre 2005 e 2015.



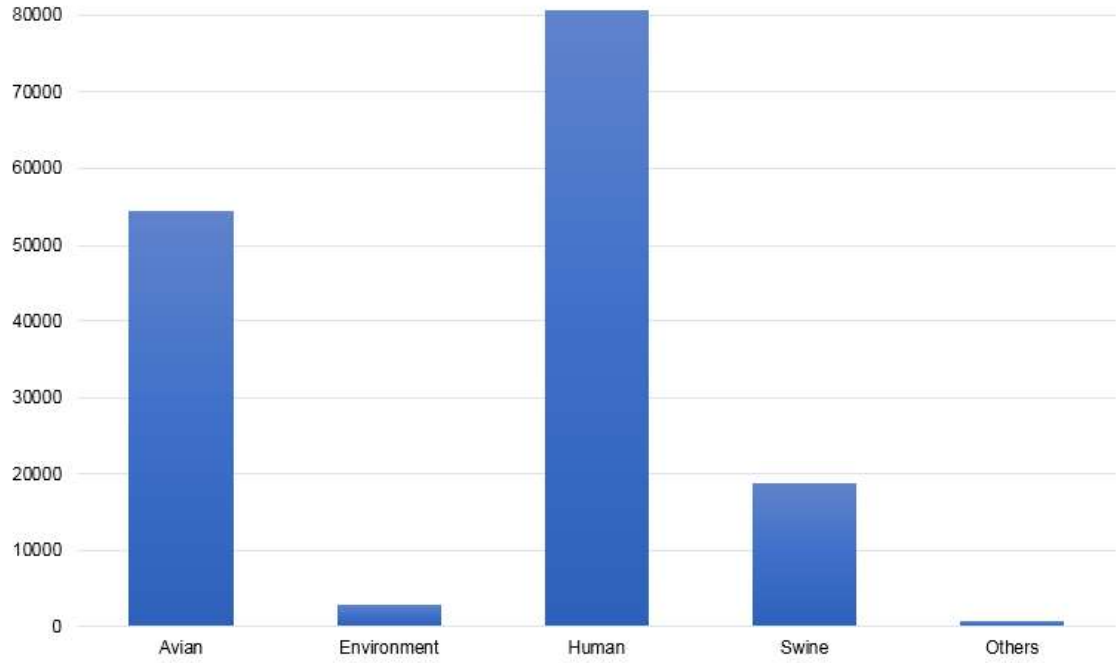
Fonte: produção do próprio autor.

Gráfico 5 – Distribuição de dados por país entre 2005 e 2015.



Fonte: produção do próprio autor.

Gráfico 6 – Distribuição de dados por hospedeiro entre 2005 e 2015.



Fonte: produção do próprio autor.

3.2 Mineração de Dados e Pós Processamento

Para aplicação da mineração de dados, foi utilizado o *software* WEKA. Foi utilizado um filtro para transformar o atributo fragmento em nominal pois o mesmo representa os oito possíveis fragmentos do RNA viral. Sendo o objetivo da mineração pesquisar relacionamentos interessantes entre os atributos do banco de dados, inicialmente foi realizada a tarefa associativa com o algoritmo Apriori, utilizando suporte de 0,1 e métrica lift. A Figura 3 apresenta as regras obtidas após a mineração de dados.

Figura 3 – Regras de associação geradas pelo algoritmo Apriori.

```
Best rules found:

 1. Subtipo=H1N1 52668 ==> Fonte=Human Ano=2009 25396    conf:(0.48) < lift:(2.65)> lev:(0.1) [15820] conv:(1.58)
 2. Fonte=Human Ano=2009 28660 ==> Subtipo=H1N1 25396    conf:(0.89) < lift:(2.65)> lev:(0.1) [15820] conv:(5.85)
 3. Fonte=Human Subtipo=H1N1 44174 ==> Ano=2009 25396    conf:(0.57) < lift:(2.36)> lev:(0.09) [14650] conv:(1.78)
 4. Ano=2009 38345 ==> Fonte=Human Subtipo=H1N1 25396    conf:(0.66) < lift:(2.36)> lev:(0.09) [14650] conv:(2.13)
 5. Subtipo=H1N1 52668 ==> Ano=2009 26500    conf:(0.5) < lift:(2.07)> lev:(0.09) [13688] conv:(1.52)
 6. Ano=2009 38345 ==> Subtipo=H1N1 26500    conf:(0.69) < lift:(2.07)> lev:(0.09) [13688] conv:(2.16)
 7. Fonte=Human 80659 ==> Subtipo=H1N1 Ano=2009 25396    conf:(0.31) < lift:(1.87)> lev:(0.08) [11836] conv:(1.21)
 8. Subtipo=H1N1 Ano=2009 26500 ==> Fonte=Human 25396    conf:(0.96) < lift:(1.87)> lev:(0.08) [11836] conv:(11.71)
 9. Fonte=Human 80659 ==> Subtipo=H1N1 44174    conf:(0.55) < lift:(1.64)> lev:(0.11) [17225] conv:(1.47)
10. Subtipo=H1N1 52668 ==> Fonte=Human 44174    conf:(0.84) < lift:(1.64)> lev:(0.11) [17225] conv:(3.03)
```

Fonte: produção do próprio autor.

Após a mineração, foram selecionadas as duas regras mais interessantes. Estas regras foram destacadas pelos seus altos valores de confiança, *lift* e convicção e estão listadas na Tabela 6. Os valores de confiança são altos, aproximando-se à unidade, indicando que as regras são fortes. Os valores de *lift* indicam dependência positiva entre os itens. Os valores muito elevados de convicção demonstram a forte dependência entre os atributos.

Tabela 6 – Regras de associação mais interessantes.

Regras	confiança	*lift*	*leverage*	convicção
2 - SE Fonte=Human E Ano=2009 ENTÃO Subtipo=H1N1	0,89	2,65	0,1	5,85
8 - SE Subtipo=H1N1 E Ano=2009 ENTÃO Fonte=Human	0,96	1,87	0,08	11,71

Para a realização da primeira tarefa de classificação, foram removidos os dados de subtipos com exceção dos subtipos H1N1 e H3N2, por serem os mais prevalentes. Além disso, foram removidos os atributos hospedeiro, local e data. Várias análises foram feitas com os atributos fragmento, tamanho e subtipo. O atributo subtipo foi selecionado para classificação, e foram testados vários algoritmos para verificar qual apresentava melhor resultado de classificação, conforme a Tabela 7.

Tabela 7 – Seleção do algoritmo para a tarefa de classificação.

Algoritmo classificador	Instâncias classificadas corretamente (%)
rules.ZeroR	54,82
bayes.NaiveBayes	54,82
functions.SimpleLogistic	54,82
trees.J48	72,50
trees.RandomTree	72,59
trees.RandomForest	72,61

O algoritmo RandomForest apresentou melhor resultado na classificação dos subtipos em sua configuração padrão (*default*) e por isso foi selecionado para mineração dos dados. Vários testes foram realizados para aumentar a qualidade dos resultados e definir o melhor modelo de classificação para os subtipos de IAV, conforme Tabela 8.

Tabela 8 – Seleção da melhor configuração do algoritmo RandomForest para modelo de classificação de subtipos de IAV (dados de 2005 a 2015).

Metodologia de teste	I (*numTrees*)*	S (*seed*)**	Instâncias classificadas corretamente (%)
*cross-validation**** 10 *fold*	100	1	72,608
*cross-validation**** 10 *fold*	100	2	72,594
*cross-validation**** 10 *fold*	100	10	72,613
*cross-validation**** 10 *fold*	100	100	72,606
*cross-validation**** 10 *fold*	200	10	72,616
*cross-validation**** 15 *fold*	200	10	72,622
*cross-validation**** 30 *fold*	200	10	72,622
*cross-validation**** 30 *fold*	200	100	72,617
*cross-validation**** 30 *fold*	2000	10	72,622
*cross-validation**** 30 *fold*	2000	1000	72,618
*cross-validation**** 5 *fold*	2000	100	72,578
*percentage split***** 66%	200	10	72,558
*percentage split***** 25%	200	10	72,174
*percentage split***** 50%	200	10	72,471
*percentage split***** 80%	200	10	72,599
*percentage split***** 99%	200	10	71,384
*training set******	200	10	72,857
*training set******	2000	10	72,857
*training set******	2000	100	72,857
*training set******	20000	1000	72,857

*I (*numTrees*) = número de árvores a serem geradas na floresta. **S (*seed*) = número randômico de sementes utilizadas para construir cada árvore. ***cross-validation* = teste com validação cruzada. ****percentage split* = utiliza uma porcentagem dos dados para teste. *****training set* = utiliza casos de treino como de teste.

Após a seleção da melhor configuração do algoritmo RandomForest, foi realizada a mineração dos dados e análise dos resultados. O resultado da mineração pode ser visualizado na matriz de classificação de subtipos de IAV apresentada na Tabela 9.

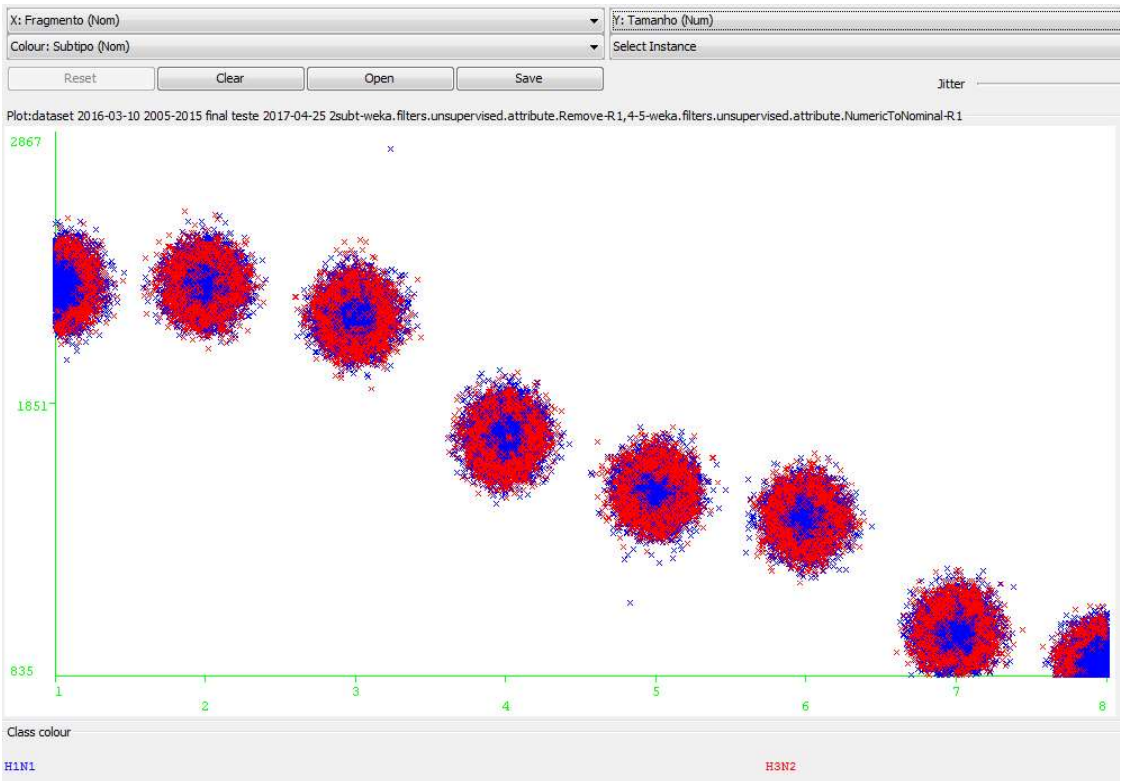
Tabela 9 – Matriz de classificação de subtipos de IAV (H1N1 X H3N2), dados de 2005 a 2015.

		Predito	
		H1N1	H3N2
Verdadeiro	H1N1	33071	19597
	H3N2	6480	36926

Os resultados mostram que o algoritmo foi capaz de classificar corretamente 72,8573% dos dados. Além disso, o algoritmo classificou corretamente 85,07% das amostras de H3N2 e 62,79% dos dados de H1N1. A menor capacidade de identificar os dados do subtipo H1N1 decorre, provavelmente, da mistura de cepas, visto que os dados contêm amostras de H1N1 sazonal do período entre 2005 e 2008 juntamente com as amostras de H1N1pdm09, que surgiu em 2009, tornando-se prevalente sobre o H1N1 sazonal.

O tamanho de cada um dos oito fragmentos do genoma do IAV possui um padrão homogêneo, com tamanho médio descrito na literatura apresentado na Tabela 1. A Figura 4 ilustra a distribuição do tamanho dos oito fragmentos dos subtipos H1N1 e H3N2 de IAV presentes na base de dados GenBank, isolados e sequenciados no período de 2005 a 2015. A tarefa de classificação com o algoritmo RandomForest resultou na identificação correta de 72,86% das amostras da base de dados e identificou 85% das cepas de H3N2 baseando-se nos tamanhos de cada um dos oito fragmentos do genoma do IAV.

Figura 4 – Tamanho dos oito fragmentos dos subtipos H1N1 e H3N2 entre 2005 e 2015.



Fonte: produção do próprio autor.

Os dados de IAV H1N1 são heterogêneos devido à mistura do IAV H1N1 sazonal com o H1N1pdm09. Por isso, foram removidos os dados de 2005 a 2008 e uma nova mineração foi realizada, a fim de observar alguma alteração nos resultados obtidos com a mineração anterior.

Após a seleção da melhor configuração do algoritmo RandomForest (Tabela 10), foi realizada a mineração dos dados e análise dos resultados. O resultado da mineração pode ser visualizado na matriz de classificação de subtipos de IAV apresentada na Tabela 11.

Esta análise apresentou melhora nos resultados de classificação. Com os dados de 2009 a 2015, o algoritmo RandomForest resultou na identificação correta de 74,37% das amostras da base de dados e aumentou para 88,15% a identificação correta das cepas de H3N2. A classificação do subtipo H1N1 aumentou para 63,6%.

Tabela 10 – Seleção da melhor configuração do algoritmo RandomForest para modelo de classificação de subtipos de IAV (dados de 2009 a 2015).

Metodologia de teste	I (*numTrees*)*	S (*seed*)**	Instâncias classificadas corretamente (%)
*cross-validation**** 15 *fold*	200	10	74,169
*percentage split***** 80%	200	10	74,190
*training set******	200	100	74,371
*training set******	2000	100	74,371

*I (*numTrees*) = número de árvores a serem geradas na floresta. **S (*seed*) = número randômico de sementes utilizadas para construir cada árvore. ****cross-validation* = teste com validação cruzada. *****percentage split* = utiliza uma porcentagem dos dados para teste. ******training set* = utiliza casos de treino como de teste.

Tabela 11 – Matriz de classificação de subtipos de IAV (H1N1pdm09 X H3N2), dados de 2009 a 2015.

		Predito	
		H1N1pdm09	H3N2
Verdadeiro	H1N1pdm09	29518	16894
	H3N2	4296	31972

Para a realização da segunda tarefa de classificação, o arquivo original resultante do pré-processamento foi editado e foram removidos os dados de hospedeiros (fontes) com exceção de aviário, humano e suíno, visto que esses são os mais comuns. Além disso, foram removidos os atributos subtipo, local e data. Várias análises foram feitas com os atributos fragmento, tamanho e hospedeiro. O atributo hospedeiro foi selecionado para classificação.

O algoritmo RandomForest foi selecionado para mineração dos dados. O melhor modelo de classificação para os hospedeiros de IAV foi definido conforme a Tabela 12.

Tabela 12 – Seleção da melhor configuração do algoritmo RandomForest para modelo de classificação de hospedeiros de IAV (Aviário/Humano/Suíno).

Metodologia de teste	I (*numTrees*)*	S (*seed*)**	Instâncias classificadas corretamente (%)
*cross-validation**** 15 *fold*	200	10	68,199
*training set*****	2000	100	68,430

*I (*numTrees*) = número de árvores a serem geradas na floresta. **S (*seed*) = número randômico de sementes utilizadas para construir cada árvore. ****cross-validation* = teste com validação cruzada. *****training set* = utiliza casos de treino como de teste.

Na seleção de dados contendo apenas três tipos de hospedeiros (aviário, humano, suíno), o algoritmo RandomForest foi capaz de identificar corretamente 68,73% das amostras, conforme matriz de classificação de hospedeiros de IAV (Tabela 13). A mineração classificou corretamente 86,24% das amostras isoladas de humanos, 62,54% das amostras aviárias e apenas 9,08% das amostras suínas. Dessas, 72,23% foram classificadas como humanas, indicando alta similaridade entre os tamanhos de fragmentos dos genomas dos subtipos que afetam humanos e suínos.

Tabela 13 – Matriz de classificação de hospedeiros de IAV (Aviário X Humano X Suíno).

		Predito		
		Aviário	Humano	Suíno
Verdadeiro	Aviário	34116	19973	462
	Humano	10323	69559	777
	Suíno	3511	13572	1706

A seguir, foram removidos os dados de hospedeiro suíno e uma nova mineração foi realizada. O modelo de classificação foi selecionado conforme resultados apresentados na Tabela 14.

Tabela 14 – Seleção da melhor configuração do algoritmo RandomForest para modelo de classificação de hospedeiros de IAV (Aviário/Humano).

Metodologia de teste	I (*numTrees*)*	S (*seed*)**	Instâncias classificadas corretamente (%)
*cross-validation**** 15 *fold*	200	10	77,017
*training set*****	200	100	77,328

*I (*numTrees*) = número de árvores a serem geradas na floresta. **S (*seed*) = número randômico de sementes utilizadas para construir cada árvore. ****cross-validation* = teste com validação cruzada. *****training set* = utiliza casos de treino como de teste.

Os resultados de classificação se mantiveram semelhantes. Nesta análise, 87,11% de amostras humanas foram identificadas corretamente, assim como 62,86% das amostras aviárias. Entretanto, com a redução dos dados, o resultado geral aumentou, e 77,33% das amostras foram classificadas corretamente conforme a matriz de classificação de hospedeiros de IAV apresentada na Tabela 15.

Tabela 15 – Matriz de classificação de hospedeiros de IAV (Aviário X Humano).

		Predito	
		Aviário	Humano
Verdadeiro	Aviário	34289	20262
	Humano	10393	70266

4 DISCUSSÃO

4.1 Limitações

Mesmo constituindo-se em uma fonte rica de análise de dados, o uso de bases de dados públicas pode resultar em erros de interpretação e outras inconsistências. Por serem dados públicos com alimentação feita por pesquisadores em todo o mundo, podem existir sequências biológicas incompletas, com erros de sequenciamento (metodologia), erros de registros de data e outros dados (subtipo, etc.).

Além disso, países mais desenvolvidos e com mais progresso científico possuem mais capacidade de coletar amostras de diferentes hospedeiros e sequenciar os genomas. Por este motivo, o número de dados dos vírus circulantes no Hemisfério Norte é muito superior ao daqueles do Hemisfério Sul. Pode-se observar, por exemplo, muitos dados provenientes dos Estados Unidos e da China, e volumes menores de outros países. Algumas razões para isto são de que nos EUA estão os principais centros de estudos genômicos de vírus influenza, e na China, pois a maioria das novas cepas de IAV originam-se no sudeste asiático[89,90].

Outro fator que influencia na quantidade de dados gerados é o fato de que a gripe é uma patologia subnotificada. No Brasil, por exemplo, a coleta de amostras de pacientes só é realizada quando o paciente apresenta Síndrome Gripal ou Síndrome Gripal Aguda Grave[91-93], e muitas pessoas que apresentam estados gripais nem mesmo procuram atendimento médico, o que resulta, por sua vez, em estatísticas de incidência incompletas. Da mesma forma, a vigilância do influenza em animais não é comum, e por isso os dados encontrados nas bases de dados públicas também não são condizentes com a realidade.

Os dados apresentados são muito heterogêneos, tendo sido sequenciadas quantidades expressivas de cepas de H1N1 e poucas ou nenhuma cepa de outros subtipos durante o mesmo período de tempo. A

distribuição de dados por hospedeiro também apresenta grandes variações, sendo mais comuns as amostras de hospedeiros humanos. Além do mais, as alterações climáticas de cada região também influenciam na quantidade de amostras obtidas para isolamento e posterior inclusão nas bases públicas.

Mesmo com uma extensa quantidade de dados, as informações contidas neste *dataset* são muito limitadas. Dados como sexo e idade do hospedeiro foram removidos por estarem presentes em uma quantidade muito pequena de registros. Além disso, este banco de dados não diferencia cepas pandêmicas como H1N1pdm09, por exemplo, de cepas sazonais.

No que se refere ao *software* Weka, a quantidade muito grande de instâncias – com muita heterogeneidade nos dados – faz com que o programa gere árvores de decisão muito grandes e muito ramificadas, de difícil visualização e interpretação. Assim, os dados precisaram ser compilados em grupos menores para que o Weka conseguisse processar.

Além disso, o *software* não possui, até o momento, algoritmos capazes de processar bases que contenham apenas sequências biológicas longas, tais como as sequências de nucleotídeos dos genomas ou sequências de aminoácidos dos proteomas. Esta limitação foi responsável pela mudança da questão de pesquisa desta dissertação de mestrado. Para a análise de genomas e proteomas, seria necessário a utilização de outras ferramentas no pré-processamento e transformação dos dados para permitir a realização das análises com o Weka. Por exemplo, a transformação de sequências genômicas e/ou proteômicas em vetores, utilizando programas em linguagem *python*, poderia permitir que a mineração de dados fosse utilizada para comparar sequências de nucleotídeos e/ou aminoácidos.

4.2 Discussão geral

Enquanto a análise estatística tradicional enfatiza a inferência nos resultados, o aprendizado de máquina enfatiza a predição. Ao se fazer uma análise estatística tradicional, o objetivo é inferir o processo pelo o qual os dados existentes foram gerados. Com a mineração de dados, é possível prever

o comportamento dos dados gerados no futuro. Desta forma, as ferramentas de mineração de dados se apresentam como alternativas na análise de bases de dados públicas, com grande potencial de geração de novos conhecimentos.

Não obstante a todas as limitações mencionadas previamente, a quantidade de dados disponível foi excessiva e compensou estas deficiências. Mesmo com uma redução de quase 75 mil dados, o algoritmo de associação utilizado foi capaz de identificar a cepa H1N1 pandêmica, mostrando forte associação entre o subtipo H1N1pdm09 e o hospedeiro humano no ano de 2009. Além disso, os algoritmos de classificação se mostraram eficientes na identificação das cepas mais prevalentes de IAV e no reconhecimento dos hospedeiros mais comuns.

Para uma adequada utilização, faz-se necessário que os modelos de dados das bases públicas sejam definidos adequadamente. Com uma estruturação das necessidades de processamento, das análises e controle semântico dos dados e com a integração das bases, será possível utilizar as mesmas na prática, evitando a perda de dados durante o processo.

O controle das pandemias de gripe requer duas etapas: a detecção precoce de variantes pandêmicas e o rápido desenvolvimento de vacinas. Para isto, é necessário um sistema de detecção rápida de cepas virais, identificação da origem e classificação de cepas de diferentes subtipos e de diferentes hospedeiros. Não obstante, a celeridade na identificação quando uma nova cepa surge com a capacidade de cruzar a barreira das espécies hospedeiras é imprescindível para reduzir o risco de novas pandemias, assim como a identificação dos fatores determinantes no tropismo do hospedeiro. Nesse sentido, a compreensão e a determinação do tropismo do hospedeiro são importantes na identificação de cepas zoonóticas do IAV capazes de atravessar a barreira das espécies e infectar seres humanos.

5 ARTIGO

COMPUTATIONAL MODELS FOR PREDICTION OF INFLUENZA A SUBTYPES AND HOST TROPISM

(Este manuscrito será submetido à revista *International Journal of Medical Informatics* conforme normas de publicação disponíveis no ANEXO B)

Fernanda Corte Real Correa[a], Ana Beatriz Gorini da Veiga[b], Silvio Cesar Cazella[c]

[a] Graduate Program in Health Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre, Brazil. fernandacr@ufcspa.edu.br

[b] Department of Basic Health Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre, Brazil. anabgv@ufcspa.edu.br

[c] Department of Exact and Applied Social Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre, Brazil. silvioc@ufcspa.edu.br

Corresponding author:

Silvio Cesar Cazella

silvioc@ufcspa.edu.br

Universidade Federal de Ciências da Saúde de Porto Alegre

Sarmento Leite 245, Porto Alegre 90050-170, RS, Brazil

ABSTRACT

Background: Influenza virus causes respiratory disease and global epidemic outbreaks every year. Predicting influenza A subtypes and host tropism is important for surveillance and control of epidemics and in the development of vaccines and new antivirals.

Objective: To develop computer models to predict influenza A subtypes and host tropism.

Methods: A large influenza dataset from GenBank including sequences from 2005 to 2015 was used to develop predictive models using different variables. Virus subtype, genome fragment number and fragment size were used for predicting subtype. Host, genome fragment number and fragment size were used for predicting host tropism. The models were built using the Random Forest algorithm available in Weka software.

Results: The predictive computer models performance in terms of accuracy rates ranged from 63% to 88% for Random Forest algorithm. The classifiers were able to correctly predict 74% of the samples of H1N1pdm09 (64%) and H3N2 (88%), and also to correctly classify the host in 77% of avian (63%) and human (87%) instances.

Conclusions: The prediction models were constructed with an influenza A virus database available openly online and were able to adequately classify the two most prevalent subtypes of influenza A as well as to differentiate between the two most common hosts of the virus. This study can contribute to future development of computational methods capable of predicting new mutant variants of influenza virus and thus preventing the spread of global outbreaks. Additionally, this study can lead to improved models for efficient and early prediction of interspecies transmission of influenza A virus.

Keywords: Influenza A virus. Data mining. Epidemic prediction. Machine learning.

BODY OF THE MANUSCRIPT

1. Introduction

Influenza is a highly contagious virus responsible for acute respiratory disease of global importance that has caused epidemics and pandemics in human population for centuries. Influenza A virus (IAV) is a negative-sense RNA virus with a segmented genome **[1]**. The eight fragments contain the genes of the hemagglutinin (HA) and neuraminidase (NA) surface protein genes, as well as other genes **[2]**.Table 1 shows the genomic structure of IAV. The virus can be further segregated into different strains according to antigenic composition of HA and NA proteins. So far, there have been described 18 variants of HA and 11 variants of NA. Current subtypes of IAV found in humans are IAV H1N1pdm09 and H3N2 **[3]**.

Table 1 – Genomic structure of influenza A virus.

genome segment	viral RNA size*	mRNA size*	codified proteins
1	2341	2320	PB2
2	2341	2320	PB1, PB1-F2
3	2233	2210	PA
4	1778	1756	HA
5	1565	1539	NP
6	1413	1391	NA
7	1027	1004, 314, 275	M1, M2
8	890	868, 396	NS1, NS2

*RNA size may differ among the virus subtypes, especially segments 4, 6 and 8. Source: adapted from Wright, Neumann e Kawaoka **[4]**.

In addition to annual influenza epidemics, IAV was the agent of four pandemics in human population **[5]**. The best-known IAV pandemic was the Spanish Flu which occurred between 1918 and 1920 and caused an estimated 50 to 100 million deaths **[6]**. The most recent pandemic started in 2009 and resulted in

more than 18 thousand deaths in more than 200 countries **[7]**. IAV pandemics result from the insertion of new influenza subtypes with antigenic composition unknown to the human population immune system. After the 2009 pandemic, seasonal H1N1 viruses were completely replaced with H1N1pdm09 viruses **[8]**.
Machine learning algorithms have been used to analyze epidemiological data. Processing large datasets can be useful in the construction of computational models for predicting epidemics, host tropism, and vaccination outcome. Studies were carried out to predict epidemics of dengue and malaria **[9,10]**.
An integrated classification and association rule mining algorithm was used to uncover the factors associated with the transformation of non-pandemic sequences into pandemic sequences, and subsequently to develop an efficient system for predicting influenza epidemics **[11]**. In a different study, ARIMA and Random Forest algorithms were applied in time series models to retrospectively analyze incidence data of avian H5N1 outbreaks in Egypt. Random Forest outperformed ARIMA in predictive ability, and researchers concluded that the algorithm is effective in predicting H5N1 epidemics in Egypt **[12]**. Random Forest was also used to build computational models for 11 influenza proteins for the prediction of host tropism. The results were highly accurate prediction models capable of determining the host tropism of individual influenza proteins **[13]**. Neural network algorithms have been used to develop computer models that can predict the results of influenza vaccination campaign, and researchers have shown that it is possible to develop useful models for predicting response to vaccination through appropriate selection of attributes **[14,15]**. Furthermore, classification models were developed to evaluate the activity of possible neuraminidase inhibitors, using Support Vector Machine and Naïve Bayesian algorithms, resulting in the discovery of nine novel neuraminidase inhibitors **[16]**.
In our study, we developed models to predict IAV subtypes and host tropism. A large GenBank database was used to build classifiers of IAV most prevalent subtypes H1N1pdm09 and H3N2 and to determine avian or human tropism of the strains. This could be significant in providing an early insight of the introduction of unique strains capable of crossing species barrier and start new influenza pandemics.

2. Material and methods

2.1. Dataset

We used the publicly available GenBank influenza dataset **[17]**. This data is collected on an ongoing basis from various countries. The dataset contains data on influenza strains isolated, sequenced, and registered by February 5, 2016. There were 232,505 instances including strains from 1902 to 2016. Incomplete and duplicated data was removed, as well as uninteresting attributes (numeric codes, host age, host sex). The data was reduced and were selected only influenza A strains. The remaining dataset contained 157,639 instances and the following attributes: host, viral genome segment, subtype, country, year, length of the sequence.

2.2. Predictor variables

The data was selected in terms of feature variables indicating the class which was being predicted. For the prediction of influenza A subtypes, the following attributes were selected: influenza A subtype, genome fragment number and fragment size. Data from 2005 to 2008 was excluded to remove seasonal H1N1. For predicting influenza A host tropism, the selection of attributes was: host, genome fragment number and fragment size. All data from 2005 to 2015 were included.

2.3. Data mining methods

We used a classification method to predict influenza A subtype and host tropism. We experimented on various classifiers in order to identify the most suited to classifying the datasets. The machine learning algorithms taken into consideration were Random Forest, Random Tree, J48, Simple Logistic, Naïve Bayes, ZeroR **[18]**. We selected Random Forest as the best algorithm to perform the data mining in the free and publicly available Weka software **[19]** for the analysis. Random Forest is an ensemble learning method for classification containing a combination of decision trees. Random trees in the forest are grown through training of a different bootstrap sample from the original data, and then by splitting leaf nodes in the trees using only a randomly selected subset of the entire feature space **[20]**.

The classifier models were conducted using a training set. In this method, the test is done on the same dataset that the classifier is trained on. The parameters of the classifiers were optimized for achieving best performance.

For each model, parameter optimization was accomplished by adjusting the number of trees and number of seeds used in the training.

We built two similar models for prediction of influenza A subtypes and host tropism. The subtype predictor contained all instances of H1N1pdm09 and H3N2 data. All data related to the other influenza A subtypes were removed from the dataset. The subtype classifier was carried out with 200 trees and 100 seeds. Other parameters were set as Weka's default values.

The host tropism prediction model contained all instances of human and avian strains data. All data related to the other influenza A hosts such as swine were removed from the dataset. The host tropism classifier was carried out with 200 trees and 100 seeds. Other parameters were set as Weka's default values.

3. Results and Discussion

The performance results of the classifiers are found in Table 2 and Table 3. The models achieved satisfactory predictive performance, as rates ranged from 64% to 88% for Random Forest algorithm. The classifiers were able to correctly predict 73% of the samples of H1N1pdm09 (63%) and H3N2 (85%), and also to correctly classify the host in 77% of avian (63%) and human (87%) instances.

Table 2 – Influenza A subtype prediction model results

classifier	instances	seeds	trees	accuracy
overall	82,680	200	100	74.37
H1N1pdm09	46,412	200	100	63,60
H3N2	36,268	200	100	88,15

Table 3 – Influenza A host prediction model results

classifier	instances	seeds	trees	accuracy
overall	135,210	200	10	77.33
human	80,659	200	10	87,11
avian	54,551	200	10	62,86

The size of the RNA genome segments of IAV has a homogeneous pattern, as described in Table 1. However, the computational models were able to correctly classify the subtypes and host tropism based only on segment number and segment size.

The use of public databases can result in mistakes and inconsistencies in the interpretation of results. Incomplete biological sequences due to sequencing errors, missing data and others can lead to miscalculation on statistical analysis. Furthermore, the records are heterogeneous. The data generated in the United States and China represent 48.70% and 12,85% of the instances, respectively. Also, due to prevalence rates, H1N1 and H3N2 characterize 33,41% and 27,54%, respectively. Host tropism is portrayed by 34,60% avian samples and 51,17% human samples.

Nonetheless, the quantity of available data was abundant and compensated these deficiencies. The classifiers were efficient and correctly predicted IAV most prevalent subtypes and host tropism.

In order to develop effective computational models for prediction of epidemiological events, it its necessary that the database used in the construction of these models have a structured and well defined format so it can be properly used avoiding loss of data in the process. The use of large global properly updated datasets can be an important factor in the development of accurate classifiers capable of predicting influenza pandemics, host tropism and other relevant information.

Epidemiological controlling of influenza pandemics requires the early detection of pandemic variants, the understanding and determination of host tropism and the rapid development of vaccines. This response demands a fast system of detection of viral strains, identification of origin and classification of subtypes and host characteristics. Therefore, the use of data mining techniques to create robust cost-effectiveness computational models capable of processing large amounts of data and generating reliable and reproducible results can become an excellent advantage in the combat of infectious diseases. The applications can vary and represent innovation in biomedical research, medical diagnostics, drugs and vaccines development and many others.

4. Conclusion

The prediction models were able to adequately classify the two most prevalent subtypes of influenza A and also to differentiate between the two most frequent hosts of the virus. The classifiers developed in this study can contribute to future models capable of predicting new variants of influenza virus and help preventing global outbreaks. Also, this study can contribute to computational models for efficient and early prediction of interspecies transmission of influenza A virus.

Author's contributions

FCRC setup the conceptual design, did most of the data analysis and interpretation and wrote the first draft and subsequent revision of this manuscript. ABGV and SCC collaborated on data analysis and in the writing and critical revision of this manuscript. All authors have read and approved the final manuscript.

Acknowledgements

The authors would like to thank Dr. Cecilia Dias Flores, Dr. Luciano Blomberg and Dr. Paulo Michel Roehe for contributing with the critical revision of this manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest statement

The authors have no competing interests to declare.

Summary points

What was already known on this topic?

- Early detection of potentially dangerous mutant strains is a crucial problem for prevention of flu pandemics and at present, there are no established methods for early identification.
- Research studies have shown that it is possible to predict mutant strains and host tropism alteration.
- However, those studies are preliminary and not applicable to prevention of global pandemics.

What this study added to our knowledge?

- We use a public database with global data for the development of predictive computational models with high accuracy rates.
- These models have the potential to be used in epidemic surveillance to prevent pandemics and selection of vaccines strain composition.
- Our study shows that early identification of potential pandemic strains and changes in host tropism is possible through machine learning tools.

References

[1] Palese P, Schulman JL. Mapping of the influenza virus genome: identification of the hemagglutinin and the neuraminidase genes. Proc Natl Acad Sci USA. 1976;73(6):2142-6.

[2] Palese P. The genes of influenza virus. Cell. 1977 Jan;10:1-10.

[3] Centers for Disease Control and Prevention. Transmission of influenza viruses from animals to people. 2014 [updated in 2017 Jan 4]. Available at: https://www.cdc.gov/flu/about/viruses/transmission.htm.

[4] Wright PF, Neumann G, Kawaoka Y. Orthomyxoviruses. In: Fields N, Knipe DM, Howley PM (Ed). Fields Virology. 5 ed. Philadelphia: Lippincott Williams & Wilkins, 2007. p.1691-1740.

[5] Palese P. Influenza: old and new threats. Nature Med. 2004;10:S82-7.

[6] Taubenberger JK, Morens DM. 1918 influenza: the mother of all pandemics. Emerg Infect Dis. 2006 Jan;12(1):15-22.
[7] World Health Organization. Pandemic (H1N1) 2009 – update 107. Emergencies preparedness, response. 2010 Jul 2 [Weekly update]. Available at: http://www.who.int/csr/don/2010_07_02/en/.
[8] World Health Organization. Influenza (seasonal). Media Centre. 2016 Nov; [Fact sheet]. [acess in 2017 Jul 3]. Available at: http://www.who.int/mediacentre/factsheets/fs211/en.
[9] Buczak AL, Koshute PT, Babin SM, Feighner BH, Lewis SH. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC Med Inform Decis Mak. 2012;12:124.
[10] Buczak AL, Baugher B, Guven E, Ramac-Thomas LC, Babin SM, et al. Fuzzy association rule mining and classification for the prediction of malaria in South Korea. BMC Med Inform Decis Mak. 2015;15:47.
[11] Kargarfard F, Sami A, Ebrahimie E. Knowledge discovery and sequence-based prediction of pandemic influenza using an integrated classification and association rule mining (CBA) algorithm. J Biomed Inform. 2015;57:181-8.
[12] Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics 2014;15:276.
[13] Eng CLP, Tong JC, Tan TW. Predicting host tropism of influenza A virus proteins using random forest. BMC Med Genomics 2014;7(Suppl 3):S1.
[14] Trtica-Majnaric L, Zekic-Susac M, Sarlija N, Vitale B. Prediction of influenza vaccination outcome by neural networks and logistic regression. J Biomed Inform 2010;43(5):774-81.
[15] Trtica-Majnaric L, Sarlija N, Vitale B. Modelling influenza vaccination outcomes. World J Vaccines. 2012;2:12-20.
[16] Lian W, Fang J, Li C, Pang X, Liu AL, Du GH. Discovery of influenza A virus neuraminidase inhibitors using support vector machine and Naïve Bayesian models. Mol Divers. 2016;20:439-51.
[17] U.S. National Library of Medicine. FTP access to GenBank data. 2016. [acess in 2016 Mar 10]. Available at: http://www.ncbi.nlm.nih.gov/genbank/ftp.

[18] Frank E, Hall MA, Witten IH. The WEKA workbench. Online appendix for "Data Mining: practical machine learning tools and techniques". San Francisco: Morgan Kaufmann, 2016.
[19] Witten IH, Frank E. Data mining: practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann, 2000.
[20] Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.

6 CONSIDERAÇÕES FINAIS E DESENVOLVIMENTO FUTURO

Os projetos genômicos em larga escala e as bases de dados públicas resultam em um número crescente de sequências biológicas e outros dados, a maioria deles sem significância fisiológica definida. A pesquisa em bioinformática contribui para o desenvolvimento de métodos para a caracterização computacional dessas sequências e para a interpretação dos dados gerados. Porém, as ferramentas tradicionais requerem muito tempo de instalação e aplicação, além de muita experiência técnica.

O vírus influenza tem muita capacidade de mutação e uma capacidade grande de mudança de hospedeiro e distribuição mundial, gerando grande variabilidade de dados. Estes bancos de dados são excelentes fontes de pesquisa, por apresentarem dados muito diversificados e em escala mundial.

Porém, para que esses bancos representem de fato uma população, vários fatores devem ser avaliados, tais como as limitações de coleta e inclusão de dados nos bancos. Além disso, e fundamental o uso de um banco de dados bem estruturado que permita o armazenamento, o acesso e o processamento de informações de forma simples e eficiente.

A mineração dos dados resultou na correta identificação do IAV H1N1pdm09. Apesar da aparente homogeneidade nos dados de fragmentos de genoma do IAV, com a mineração de dados foi possível diferenciar os dois subtipos mais prevalentes do IAV. A tarefa de classificação com o algoritmo RandomForest resultou na identificação correta do subtipo de 74% das amostras da base de dados, além de classificar 64% das amostras de H1N1 e 88% das cepas de H3N2 baseando-se nos tamanhos de cada um dos oito fragmentos do genoma do IAV. Outrossim, com a mineração de dados foi possível diferenciar o hospedeiro de 77% das amostras aviárias (63%) e humanas (87%). Assim, é possível concluir que é possível descobrir novos conhecimentos através da análise de banco de dados com informações sobre o genoma do vírus influenza através do uso de uma ferramenta de mineração de dados.

O *software* Weka se apresentou como uma ferramenta útil na classificação do subtipo de cepas de IAV através do tamanho dos fragmentos de RNA sequenciados. Outros modelos computacionais de classificação podem ser desenvolvidos para classificar outros subtipos de IAV, e também aprimorados para que sejam capazes de diferenciar cepas H1N1 prévias ao H1N1pdm09 do mesmo. Ademais, também é possível realizar novas análises comparando os dados de cada fragmento individualmente, para verificar quais fragmentos são mais preditores na classificação de subtipos e no tropismo de hospedeiros.

A mineração de dados revelou-se como uma excelente opção para análise de dados biológicos, com *softwares* mais robustos com ótimo custo-benefício, capazes de gerar grandes volumes de resultados confiáveis de maneira rápida e reprodutível. Na área da saúde, a mineração possui um potencial imensurável de aplicações, podendo ser utilizada no auxílio de pesquisas biomédicas, na indicação de diagnósticos médicos mais precisos, na seleção individualizada de tratamentos medicamentosos, na predição de epidemias, na seleção de vacinas, entre outros, e pode ser considerada uma das tecnologias mais promissoras da atualidade.

Por ter um potencial ilimitado de aplicações, a mineração de dados pode ser utilizada para analisar bases de dados públicas com uma quantidade maior de dados e com mais variáveis epidemiológicas, tais como sexo, idade, vacinação contra influenza, infecções por influenza prévias, comorbidades, etc. Além disso, também é possível fazer análises de sequências genômicas e/ou proteômicas, associando-se a mineração de dados a outras técnicas para a transformação prévia destes dados biológicos – tais como programação em linguagem *python* –, de forma que os mesmos possam ser processados pelos algoritmos de mineração.

Através da combinação de ferramentas, é possível que os dados sejam transformados de forma que possam ser utilizados na mineração de dados. Além disso, as bases de dados podem ser combinadas entre si, reunindo dados de sequências de nucleotídeos – transformadas em vetores – juntamente com outras informações epidemiológicas, possibilitando, dessa forma, a descoberta de novos conhecimentos sobre os patógenos tais como

predição de comportamentos futuros como novas mutações, o surgimento de novos subtipos, o risco de que uma cepa migre para um novo hospedeiro e até mesmo novas epidemias. Também é possível ‘utilizar ferramentas de alinhamento de sequências previamente, para ajudar a selecionar as sequências que serão transformadas.

Além disso, trabalhos futuros na construção de modelos de computação capazes de predizer o surgimento de novas cepas, predizer diretamente a transmissão de IAV interespécies, predizer resposta a campanhas de vacinação, entre outros, devem ser realizados para que as ferramentas de mineração de dados possam ser aplicadas em todo o seu potencial. A base de dados utilizada neste trabalho representa um excelente instrumento a ser aplicado no desenvolvimento de modelos computacionais para a análise do IAV.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Lima-Camara TN. Emerging arboviruses and public health challenges in Brazil. Rev Saúde Pública. 2016;50:36.

2. Goldani LZ. Yellow fever outbreak in Brazil, 2017. Braz J Infect Dis. 2017;21(2):123-4.

3. World Health Organization. Pandemic (H1N1) 2009 – update 107. Emergencies preparedness, response. 2010 Jul 2 [Weekly update]. [acesso em 2016 Apr 8]. Disponível em: http://www.who.int/csr/don/2010_07_02/en/.

4. Yang X, Yang H, Zhou G, Zhao, G. Infectious disease in the genomic era. Annu Rev Genomics Hum Genet. 2008;9:21-48.

5. He CQ, Han GZ, Wang D, Liu W, Li GR, Liu XP, et al. Homologous recombination evidence in human and swine influenza A viruses. Virology. 2008 Oct 10;380(1):12-20.

6. Nelson MI, Simonsen L, Viboud C, Miller MA, Holmes EC. Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. PLoS Pathog. 2007 Sep;3(9):e131.

7. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009 Sep;5(9):e1000520.

8. Jiawei H, Kamber M, Pei J. Data mining: concepts and techniques. 3 ed. San Francisco: Morgan Kaufmann, 2011.

9. Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM. 1996;39(11):27-34.

10. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann, 2000.

11. Fernandez-Lozano C, Gestal M, González-Díaz H, Dorado J, Pazos A, Munteanu CR. Markov mean properties for cell death-related protein classification. J Theor Biol. 2014 May 21;349:12-21.

12. Frank SA. Immunology and evolution of infectious disease. Princeton (NJ): Princeton University Press, 2002. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK2394.

13. Hilleman MR. Realities and enigmas of human viral influenza: pathogenesis, epidemiology and control. Vaccine. 2002 Aug 19;20(25-26):3068-98.

14. Chen R, Holmes ED. The evolutionary dynamics of human influenza B virus. J Mol Evol. 2008 Jun;66(6):655-63.

15. Vedovello D. Diversidade genética dos vírus influenza A, detectados em crianças de São Paulo [tese]. São Paulo: Universidade de São Paulo; 2011.

16. Wright PF, Neumann G, Kawaoka Y. Orthomyxoviruses. In: Fields N, Knipe DM, Howley PM (Ed). Fields Virology. 5 ed. Philadelphia: Lippincott Williams & Wilkins, 2007. p.1691-1740.

17. Mueller M, Renzullo S, Brooks R, Ruggli N, Hofmann MA. Antigenic characterization of recombinant hemagglutinin proteins derived from different avian influenza virus subtypes. PLoS One. 2010 Feb 5;5(2):e9097.

18. Centers for Disease Control and Prevention. Transmission of influenza viruses from animals to people. 2014 [atualizada em 2017 Jan 4; acesso em 2017 Feb 2]. Disponível em: https://www.cdc.gov/flu/about/viruses/transmission.htm.

19. World Health Organization. Reconsideration of influenza A virus nomenclature: a WHO Memorandum. Bull World Health Organ. 1979;57(2):227-33.

20. World Health Organization. A revision of the system of nomenclature for influenza viruses: a WHO Memorandum. Bull World Health Organ. 1980;58(4):585-91.

21. Shope RE. Swine influenza. I. Experimental transmission and pathology. J Exp Med. 1931 Jul 31;54(3):349-59.

22. Smith TF, Burgert EO Jr, Dowdle WR, Noble GR, Campbell RJ, Van Scoy RE. Isolation of swine influenza virus from autopsy lung tissue of man. N Engl J Med. 1976 Mar 25;294(13):708-10.

23. Palese P, Schulman JL. Mapping of the influenza virus genome: identification of the hemagglutinin and the neuraminidase genes. Proc Natl Acad Sci USA. 1976;73(6):2142-6.

24. Palese P. The genes of influenza virus. Cell. 1977 Jan;10:1-10.

25. Vasin AV, Temkina AO, Egorov VV, Klotchenko SA, Plotnikova MA, Kiselev OI. Molecular mechanisms enhancing the proteome of influenza A viruses: an overview of recently discovered proteins. Virus Res. 2014 Jun 24;185:53-63.

26. Cox NJ, Subbarao K. Influenza. Lancet. 1999 Oct 9;354(9186):1277-82.

27. Weis W, Brown JH, Cusack S, Paulson JC, Skehel JJ, Wiley DC. Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. Nature. 1988;333:426-31.

28. Samji T. Influenza A: understanding the viral life cycle. Yale J Biol Med. 2009 Dec;82(4):153-9.

29. Medina RA, Garcia-Sastre A. Influenza A viruses: new research developments. Nat Rev Microbiol. 2011;9(8):590-603.

30. Forleo-Neto E, Halker E, Santos VJ, Paiva TM, Toniolo-Neto J. Influenza. Rev Soc Bras Med Trop. 2003 Mar-Apr;36(2):267-74.

31. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, et al. Mapping the antigenic and genetic evolution of influenza virus. Science. 2004 Jul 16;305(5682):371-6.

32. Aggarwal S, Bradel-Tretheway B, Yakimoto T, Dewhusrt S, Kim B. Biochemical characterization of enzyme fidelity of influenza A virus RNA polymerase complex. PLoS One. 2010 Apr 29;5(4):e10372.

33. Kilbourne ED. Influenza pandemics of the 20th Century. Emerg Infect Dis. 2006 Jan;12(1):9-14.

34. World Health Organization. Influenza (seasonal). Media Centre. 2014 Mar;211[Fact sheet]. [acesso em 2016 Apr 8]. Disponível em: http://www.who.int/mediacentre/factsheets/fs211/en.

35. Palese P. Influenza: old and new threats. Nature Med. 2004;10:S82-7.

36. Koel BF, Mögling R, Chutinimitkul S, Fraaij PL, Burke DF, van der Vliet S, et al. Identification of amino acid substitutions supporting antigenic change of influenza A(H1N1)pdm09 viruses. J Virol. 2015 Apr;86(7):3763-75.

37. Luk J, Gross P, Thompson WW. Observations on mortality during the 1918 influenza pandemic. Clin Infect Dis. 2001 Oct 15;33(8):1375-8.

38. Taubenberger JK, Morens DM. 1918 influenza: the mother of all pandemics. Emerg Infect Dis. 2006 Jan;12(1):15-22.

39. Trilla A, Trilla G, Daer C. The 1918 “Spanish flu” in Spain. Clin Infect Dis. 2008 Sep 1;47(5):668-73.

40. Smith W, Andrewes CH, Laidlaw PP. A virus obtained from influenza patients. Lancet. 1933 Jul 8;222(5732):66-8.

41. Zhdanov VM, Ritova VV, Orlova AV, Sokolova NN. The characteristics of influenza-virus strains isolated in 1957. Lancet. 1957 Oct 12;273(6998):735-6.

42. Doraisingham S, Goh KT, Ling AE, Yu M. Influenza surveillance in Singapore: 1972-86. Bull World Health Organ. 1988;66(1):57-63.

43. Viboud C, Grais RF, Lafont BAP, Miller MA, Simonsen L. Multinational impact of the 1968 Hong Kong influenza pandemic: evidence for a smoldering pandemic. J Infect Dis. 2005 Jul 15;192(2):233-48.

44. Guan Y, Vijaykrishna D, Bahl J, Zhu H, Wang J, Smith GJ. The emergence of pandemic influenza viruses. Protein Cell. 2010 Jan 1;1(1):9-13.

45. Shinde V, Bridges CB, Uyeki TM, Shu B, Balish A, Xu X, et al. Triple-reassortant swine influenza A (H1) in humans in the United States, 2005-2009. N Engl J Med. 2009 Jun 18;360(25):2616-25.

46. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009 Jun 25;459:1122-25.

47. Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. N Engl J Med. 2009 June 18;360(25):2605-15.

48. World Health Organization. World now at the start of 2009 influenza pandemic. Media Centre. 2009 Jun 11 [Statement]. [acesso em 2016 Apr 8]. Disponível em: http://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en.

49. World Health Organization. Recommended composition of influenza virus vaccines for use in the 2017 southern hemisphere influenza season. World Health Organ. 2016 Sep 29:1-8 [atualizada em 2016 Oct 6; acesso em 2017 Feb 2]. Disponível em: http://www.who.int/influenza/vaccines/virus/recommendations/201609_recommendation.pdf?ua=1.

50. Hagen JB. The origins of bioinformatics. Nat Rev Genet. 2000 Dec;1(3):231-6.

51. Ouzounis CA. Rise and demise of bioinformatics? Promise and progress. PLoS Comput Biol. 2012;8(4):e1002487.

52. Hogeweg P. The roots of bioinformatics in theoretical biology. PLoS Comput Biol. 2011 Mar;7(3):e1002021.

53. Seibel LFB, Lemos M, Lifschitz S. Bancos de dados de genoma. Proc Braz Databases Symposium Tutorials. 2000:514-53.

54. Claverie JM. From bioinformatics to computational biology. Genome Res. 2000;10:1277-9.

55. Linoff GS, Berry MJA. Data mining techniques: for marketing, sales, and customer relationship management. 3 ed. New York: John Wiley & Sons, 2011.

56. Lee GW, Kim S. Genome data mining for everyone. BMB reports. 2008;41(11):757-64.

57. Fayyad U. Data mining. Communications of the ACM. 2003:496-9.

58. Camilo CO, da Silva JC. Mineração de dados: conceitos, tarefas, métodos e ferramentas. Technical Report RT-INF 001-09. Goiás: Universidade Federal de Goiás, Aug 2009.

59. Vasconcelos LMR, Carvalho CL. Aplicação de Regras de Associação para Mineração de Dados na Web. Technical Report RT-INF 004-04. Goiás: Universidade Federal de Goiás, Nov 2004.

60. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Twentieth international conference on very large data bases. 1994:478-99.

61. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: Fourth international conference on knowledge discovery and data mining. 1998:80-6.

62. Amo S. Técnicas de mineração de dados. Universidade Federal de Uberlândia, 2003.

63. Goldschmidt R, Passos E. Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações e aplicações. São Paulo: Elsevier, 2005.

64. John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. In: Eleventh conference on uncertainty in artificial intelligence. San Mateo, 1995.

65. Landwehr N, Hall M, Frank E. Logistic Model Trees. 2005.

66. Summer M, Frank E, Hall M. Speeding up logistic model tree induction. In: Ninth European conference on principles and practice of knowledge discovery in databases. 2005:675-83.

67. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.

68. James G, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. London: Springer, 2013.

69. Lorenzett CDC, Telöcken AV. Estudo comparative entre os algoritmos de mineração de dados Random Forest e J48 na tomada de decisão. Anais II Simp Pesq Desenv Comp. 2016;1-10.

70. Liaw A, Wiener M. Classification and regression by randomForest. R J. 2002 Dec;2(3):18-22.

71. Frank E, Hall MA, Witten IH. The WEKA workbench. Online appendix for "Data Mining: practical machine learning tools and techniques". San Francisco: Morgan Kaufmann, 2016.

72. Baxevanis AD. The importance of biological databases in biological discovery. Curr Protoc Bioinformatics. 2009 Sep;Chapter 1:Unit 1.1.

73. Catanho M, Miranda AB. Comparando genomas: bancos de dados e ferramentas computacionais para a análise comparativa de genomas procarióticos. Rev Eletr Com Inf Inov Saúde. 2007;1(2):Sup335-58.

74. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2010 Jan;38(Database issue):D46-51.

75. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2010 Jan;38(Database issue):D5-16.

76. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2012;40(Database issue):D48-53.

77. Benson DA, Cavanaugh M, Clark K, Karsch-Mirzrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2017;45(Database issue):D37-42.

78. Weiss SM, Kulikowski CA. Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. San Francisco: Morgan Kaufmann, 1991.

79. Carvalho DR, Dallagassa MR. Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas. AtoZ. 2014;3(2):82-6.

80. Buczak AL, Koshute PT, Babin SM, Feighner BH, Lewis SH. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC Med Inform Decis Mak. 2012;12:124.

81. Buczak AL, Baugher B, Guven E, Ramac-Thomas LC, Babin SM, et al. Fuzzy association rule mining and classification for the prediction of malaria in South Korea. BMC Med Inform Decis Mak. 2015;15:47.

82. Kargarfard F, Sami A, Ebrahimie E. Knowledge discovery and sequence-based prediction of pandemic influenza using an integrated classification and association rule mining (CBA) algorithm. J Biomed Inform. 2015;57:181-8.

83. Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics 2014;15:276.

84. Eng CLP, Tong JC, Tan TW. Predicting host tropism of influenza A virus proteins using random forest. BMC Med Genomics 2014;7(Suppl 3):S1.

85. Trtica-Majnaric L, Zekic-Susac M, Sarlija N, Vitale B. Prediction of influenza vaccination outcome by neural networks and logistic regression. J Biomed Inform 2010;43(5):774-81.

86. Trtica-Majnaric L, Sarlija N, Vitale B. Modelling influenza vaccination outcomes. World J Vaccines. 2012;2:12-20.

87. Lian W, Fang J, Li C, Pang X, Liu AL, Du GH. Discovery of influenza A virus neuraminidase inhibitors using support vector machine and Naïve Bayesian models. Mol Divers. 2016;20:439-51.

88. U.S. National Library of Medicine. FTP access to GenBank data. 2016. [acesso em 2016 Mar 10]. Disponível em: http://www.ncbi.nlm.nih.gov/genbank/ftp.

89. Wen F, Bedford T, Cobey S. Explaining the geographical origins of seasonal influenza A (H3N2). Proc R Soc. 2016 Sep 14;283(1838):1-9.

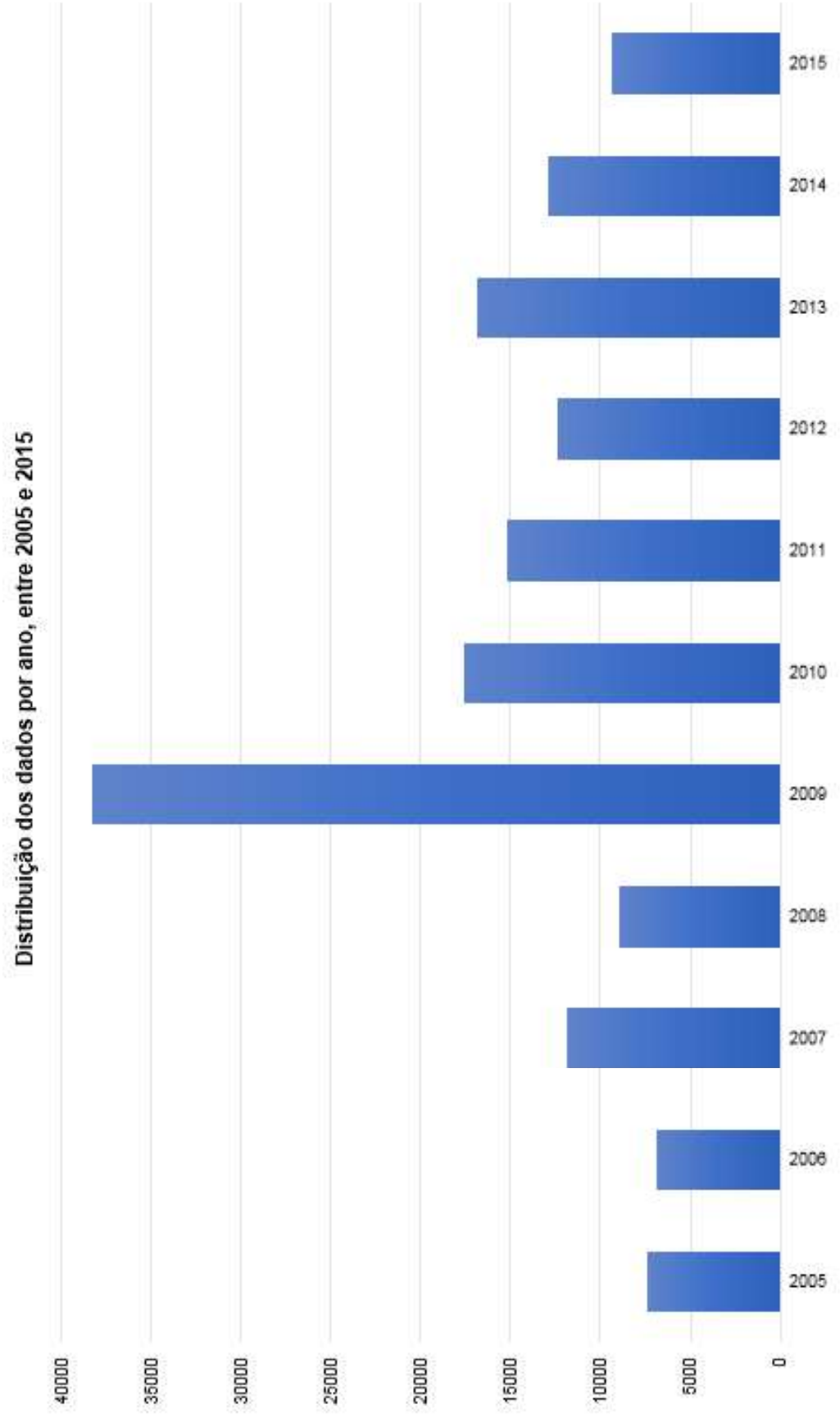
90. Chan J, Holmes A, Rabadan R. Network analysis of global influenza spread. PLoS Comput Biol. 2010 Nov 18;6(11):e1001005.

91. World Health Organization. WHO Global Influenza Surveillance Network. Manual for the laboratory diagnosis and virological surveillance of influenza. Geneva: World Health Organization, 2011. 153p.

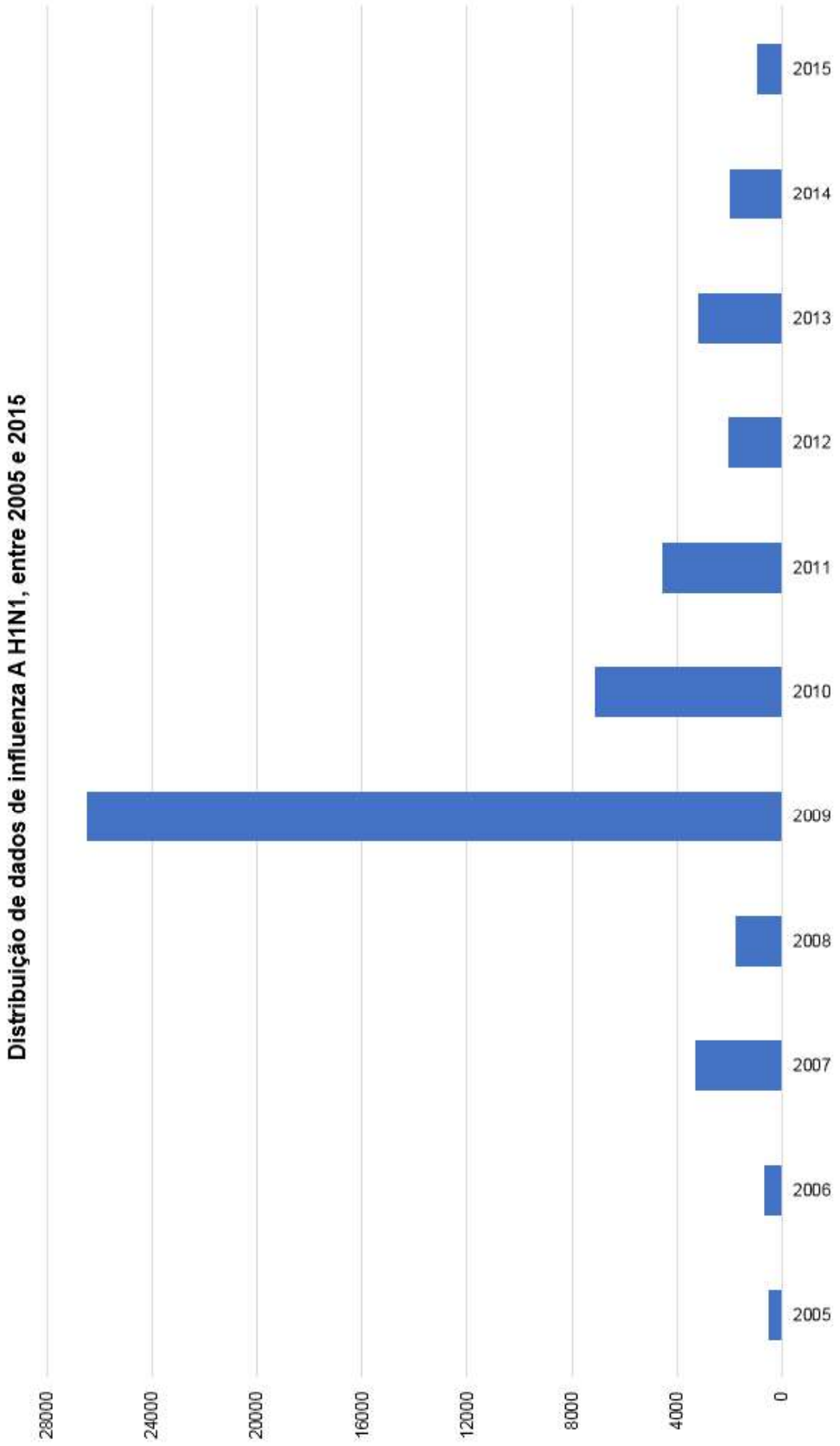
92. Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis. Guia para a Rede Laboratorial de Vigilância de Influenza no Brasil. Brasília: Ministério da Saúde, 2016. 64p.

93. Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Informe epidemiológico. Influenza: monitoramento até a semana epidemiológica 08 de 2017. 2017 [acesso em 2017 Jul 5]. Disponível em: http://portalarquivos.saude.gov.br/images/pdf/2017/marco/08/Informe%20Epidemiolgico_Influenza-2017-SE-08.pdf.

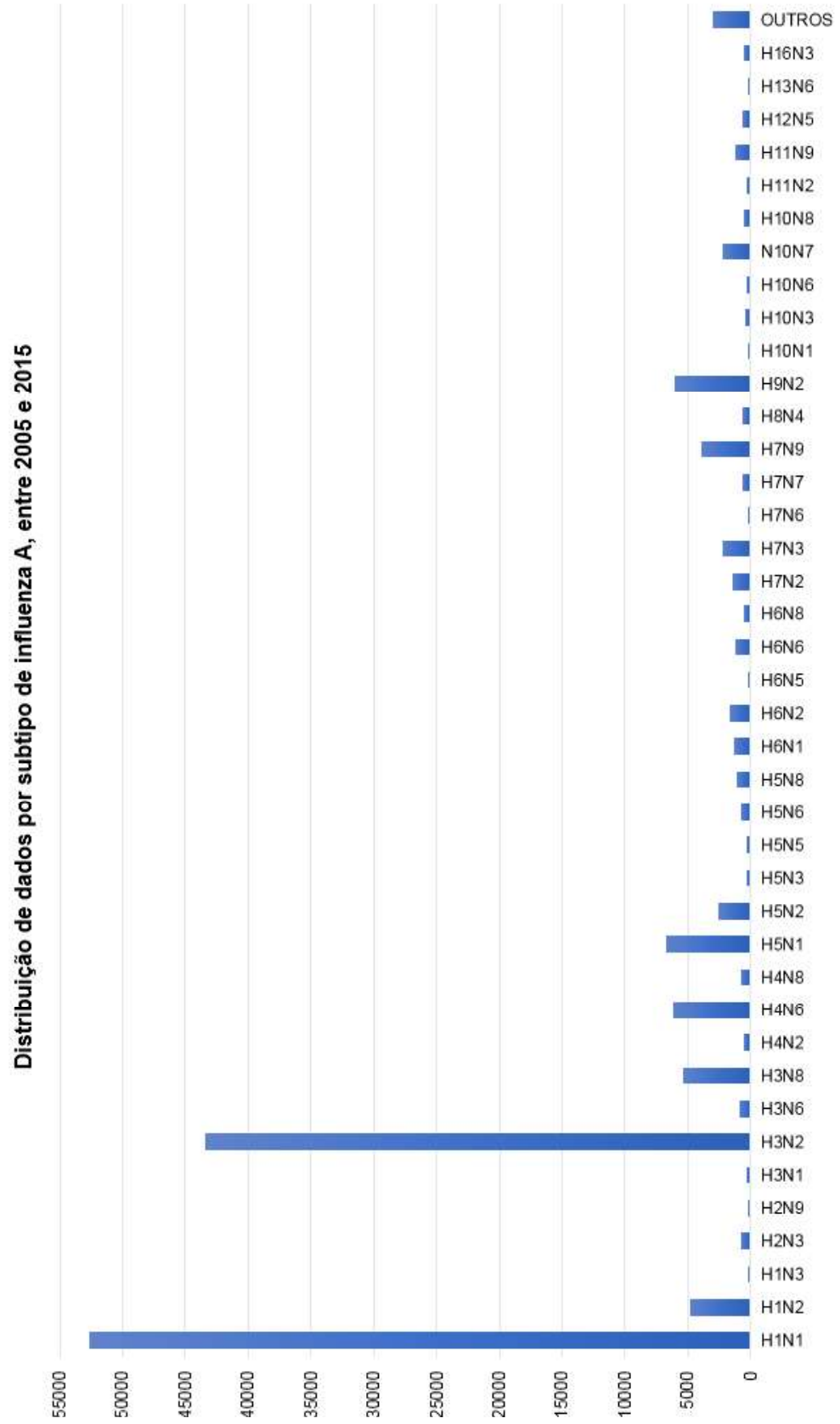
APÊNDICE A – Gráfico de distribuição de dados por ano entre 2005 e 2015.



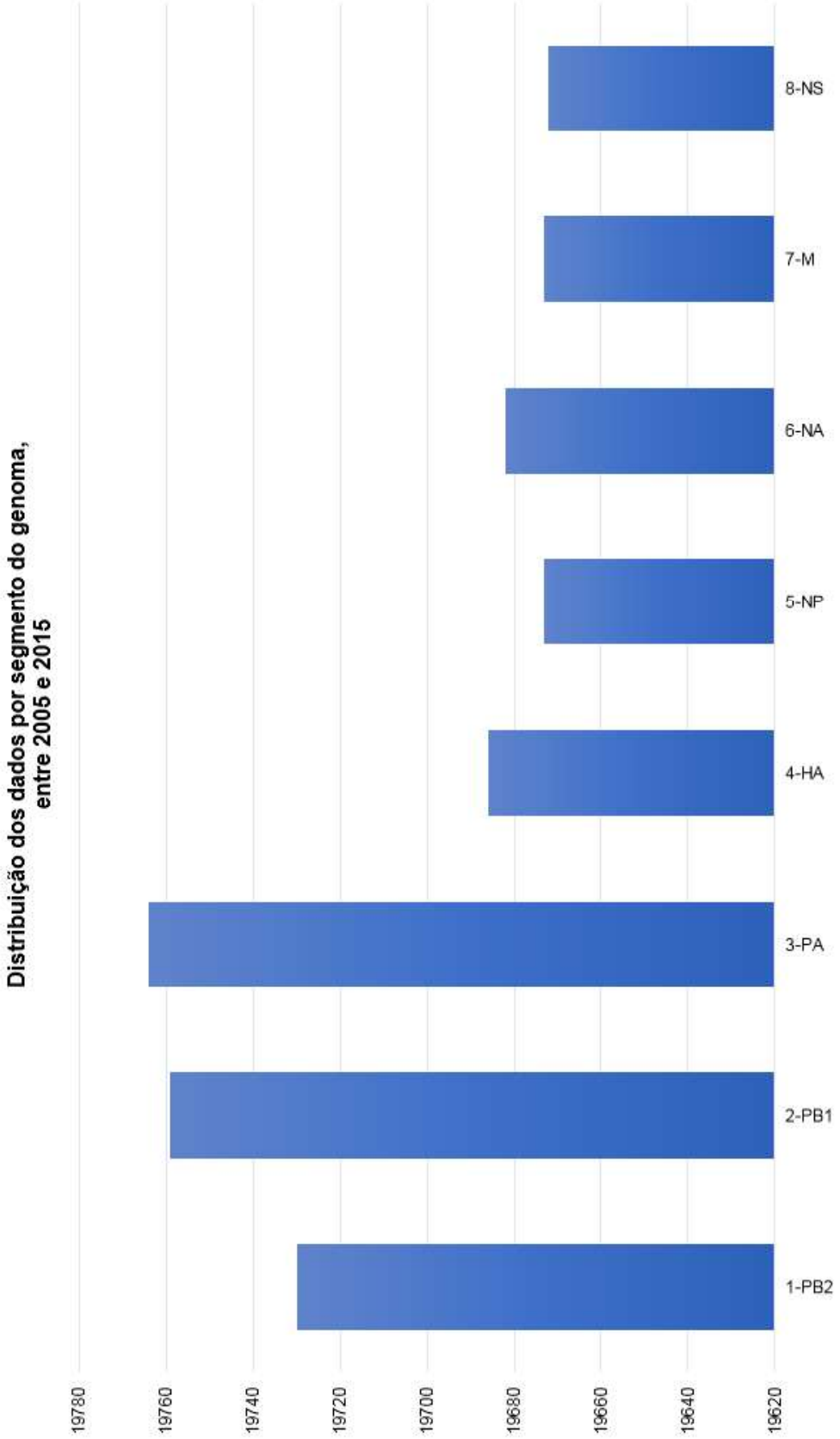
APÊNDICE B – Gráfico de distribuição de dados de influenza A H1N1 entre 2005 e 2015.



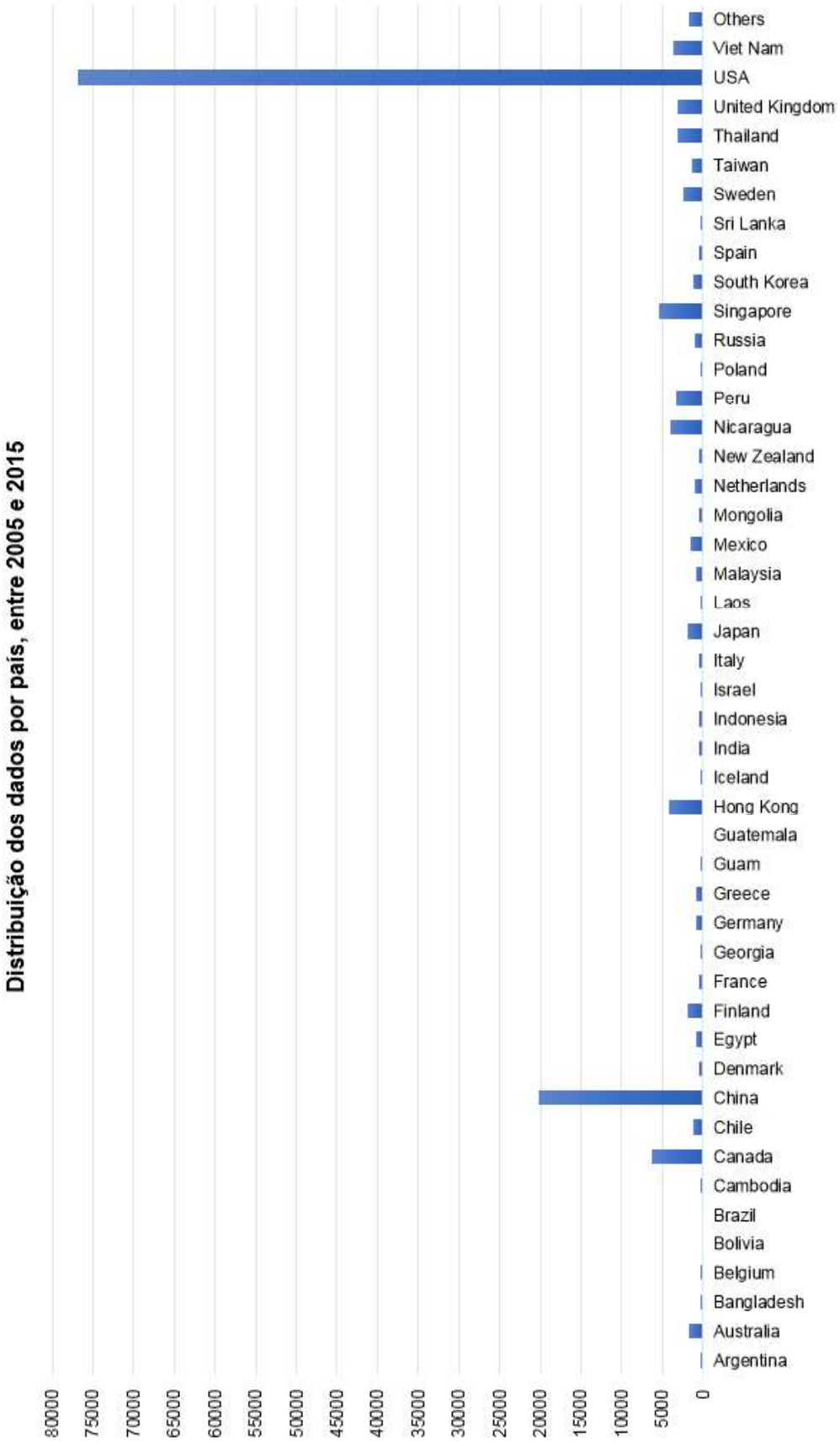
APÊNDICE C – Gráfico de distribuição de dados por subtipo entre 2005 e 2015.



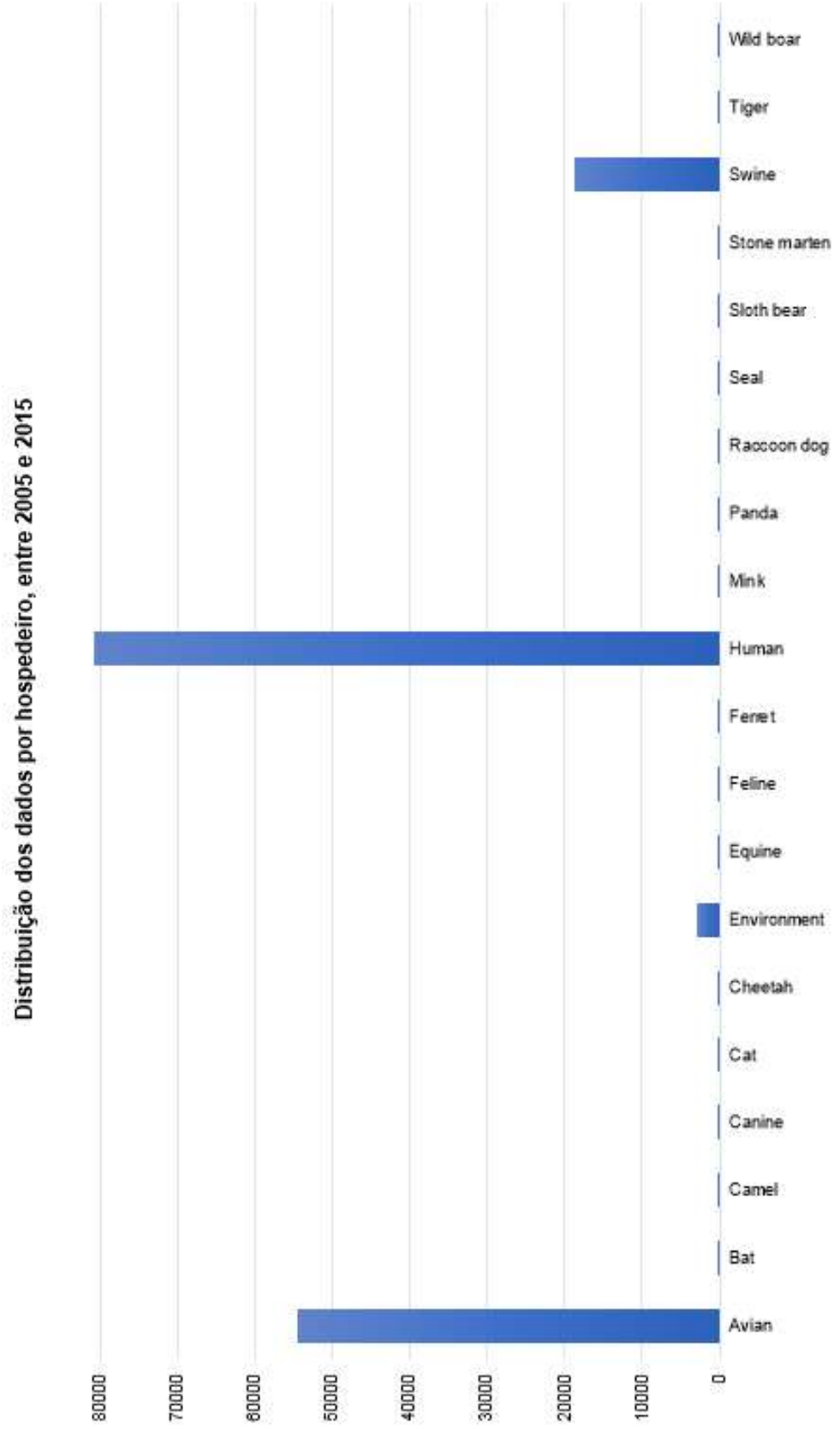
APÊNDICE D – Gráfico de distribuição de dados por fragmento de genoma entre 2005 e 2015.



APÊNDICE E – Gráfico de distribuição de dados por país entre 2005 e 2015.



APÊNDICE F – Gráfico de distribuição de dados por hospedeiro entre 2005 e 2015.



ANEXO A – Registro na Comissão de Pesquisa da UFCSPA

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA EDUCAÇÃO

UFCSPA
UNIVERSIDADE FEDERAL DE CIÊNCIAS DA SAÚDE DE PORTO ALEGRE
COMISSÃO DE PESQUISA

Certificado

Certificamos que o projeto de pesquisa intitulado *"Análise do genoma e do proteoma de vírus patogênicos humanos através da aplicação de técnicas de mineração de dados"*, de Silvio César Cazella, está registrado na Comissão de Pesquisa da Universidade Federal de Ciências da Saúde de Porto Alegre sob o número 039/2016. Salientamos que este registro **não autoriza o pesquisador a coletar ou analisar dados oriundos de sujeitos de pesquisa**.

Salientamos ainda que tal registro **não garante** a concessão de recursos financeiros por parte da UFCSPA a este projeto de pesquisa.

Porto Alegre, 30 de setembro de 2016.

Paulo Ricardo Gazzola Zen
Coordenador Geral da Pesquisa
UFCSPA

ANEXO B – Normas de publicação da revista

INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS

The International Journal of Medical Informatics is the official journal of the European Federation for Medical Informatics (EFMI) and IMIA.

AUTHOR INFORMATION PACK

TABLE OF CONTENTS

ISSN: 1386-5056

DESCRIPTION

International Journal of Medical Informatics provides an international medium for dissemination of original results and interpretative reviews concerning the field of medical informatics. The Journal emphasizes the evaluation of systems in healthcare settings.
The scope of journal covers: Information systems, including national or international registration systems, hospital information systems, departmental and/or physician's office systems, document handling systems, electronic medical record systems, standardization, systems integration etc.; Computer-aided medical decision support systems using heuristic, algorithmic and/or statistical methods as exemplified in decision theory, protocol development, artificial intelligence, etc. Educational computer based programs pertaining to medical informatics or medicine in general; Organizational, economic, social, clinical impact, ethical and cost-benefit aspects of IT applications in health care.

Short technical communications concerning (solved) problems in implementing or using existing information systems are welcome. Review articles concerning subjects falling in the scope of the journal are also invited.

AUDIENCE

Those working in computing applied to the medical and life sciences, Biomathematicians, Life Sciences Researchers, Bioengineers, Radiologists.

IMPACT FACTOR

2015: 2.363 © Thomson Reuters Journal Citation Reports 2016

GUIDE FOR AUTHORS

Your Paper Your Way

We now differentiate between the requirements for new and revised submissions. You may choose to submit your manuscript as a single Word or PDF file to be used in the refereeing process. Only when your paper is at the revision stage, will you be requested to put your paper in to a 'correct format' for acceptance and provide the items required for the publication of your article.
To find out more, please visit the Preparation section below.

Aims and Scope

International Journal of Medical Informatics provides an international medium for dissemination of original results and interpretative reviews concerning the field of medical informatics. The Journal emphasizes the evaluation of systems in healthcare settings. The scope of journal covers:
Information systems, including national or international registration systems, hospital information systems, departmental and/or physician's office systems, document handling systems, electronic medical record systems, standardization, systems integration etc.;
Computer-aided medical decision support systems using heuristic, algorithmic and/or statistical methods as exemplified in decision theory, protocol development, artificial intelligence, etc.
Educational computer based programs pertaining to medical informatics or medicine in general;
Organizational, economic, social, clinical impact, ethical and cost-benefit aspects of IT applications in health care.

Short technical communications concerning (solved) problems in implementing or using existing information systems are welcome. Review articles concerning subjects falling in the scope of the journal are also invited.

General Considerations

IJMI has adopted the guidelines of the International Committee of Medical Journal Editors (ICMJE). Some of the important issues are noted below. Visit http://www.icmje.org for more details.

Contact details

All submissions should be made through Elsevier's Editorial System (EES) via http://ees.elsevier.com/ijmi).

Submission checklist

You can use this list to carry out a final check of your submission before you send it to the journal for review. Please check the relevant section in this Guide for Authors for more details.

Ensure that the following items are present:

One author has been designated as the corresponding author with contact details:
• E-mail address
• Full postal address

All necessary files have been uploaded:
Manuscript:
• Include keywords
• All figures (include relevant captions)
• All tables (including titles, description, footnotes)
• Ensure all figure and table citations in the text match the files provided
• Indicate clearly if color should be used for any figures in print
Graphical Abstracts / Highlights files (where applicable)
Supplemental files (where applicable)

Further considerations
• Manuscript has been 'spell checked' and 'grammar checked'
• All references mentioned in the Reference List are cited in the text, and vice versa
• Permission has been obtained for use of copyrighted material from other sources (including the Internet)
• Relevant declarations of interest have been made
• Journal policies detailed in this guide have been reviewed
• Referee suggestions and contact details provided, based on journal requirements

For further information, visit our Support Center.

BEFORE YOU BEGIN

Ethics in publishing

Please see our information pages on Ethics in publishing and Ethical guidelines for journal publication.

Human and animal rights

If the work involves the use of human subjects, the author should ensure that the work described has been carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans; Uniform Requirements for manuscripts submitted to Biomedical Journals. Authors should include a statement in the manuscript that informed consent was obtained for experimentation with human subjects. The privacy rights of human subjects must always be observed.

All animal experiments should comply with the ARRIVE guidelines and should be carried out in accordance with the U.K. Animals (Scientific Procedures) Act, 1986 and associated guidelines, EU Directive 2010/63/EU for animal experiments, or the National Institutes of Health guide for the care and use of Laboratory animals (NIH Publications No. 8023, revised 1978) and the authors should clearly indicate in the manuscript that such guidelines have been followed.

Declaration of interest

All authors must disclose any financial and personal relationships with other people or organizations that could inappropriately influence (bias) their work. Examples of potential conflicts of interest include employment, consultancies, stock ownership, honoraria, paid expert testimony, patent applications/ registrations, and grants or other funding. If there are no conflicts of interest then please state this: 'Conflicts of interest: none'. More information.

Submission declaration and verification

Submission of an article implies that the work described has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see 'Multiple, redundant or concurrent publication' section of our ethics policy for more information), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. To verify originality, your article may be checked by the originality detection service CrossCheck.

Authorship

All authors should have made substantial contributions to all of the following: (1) the conception and design of the study, or acquisition of data, or analysis and interpretation of data, (2) drafting the article or revising it critically for important intellectual content, (3) final approval of the version to be submitted.

Acknowledgements

All contributors who do not meet the criteria for authorship as defined above should be listed in an acknowledgements section. Examples of those who might be acknowledged include a person who provided purely technical help, writing assistance, or a department chair who provided only general support. Authors should disclose whether they had any writing assistance and identify the entity that paid for this assistance. Financial support like grants should also be mentioned in the acknowledgements section.

Changes to authorship

Authors are expected to consider carefully the list and order of authors **before** submitting their manuscript and provide the definitive list of authors at the time of the original submission. Any addition, deletion or rearrangement of author names in the authorship list should be made only **before** the manuscript has been accepted and only if approved by the journal Editor. To request such a change, the Editor must receive the following from the **corresponding author**: (a) the reason for the change in author list and (b) written confirmation (e-mail, letter) from all authors that they agree with the addition, removal or rearrangement. In the case of addition or removal of authors, this includes confirmation from the author being added or removed.

Only in exceptional circumstances will the Editor consider the addition, deletion or rearrangement of authors **after** the manuscript has been accepted. While the Editor considers the request, publication of the manuscript will be suspended. If the manuscript has already been published in an online issue, any requests approved by the Editor will result in a corrigendum.

Clinical trial results

In line with the position of the International Committee of Medical Journal Editors, the journal will not consider results posted in the same clinical trials registry in which primary registration resides to be prior publication if the results posted are presented in the form of a brief structured (less than 500 words) abstract or table. However, divulging results in other circumstances (e.g., investors' meetings) is discouraged and may jeopardise consideration of the manuscript. Authors should fully disclose all posting in registries of results of the same or closely related work.

Reporting clinical trials

Randomized controlled trials should be presented according to the CONSORT guidelines. At manuscript submission, authors must provide the CONSORT checklist accompanied by a flow diagram that illustrates the progress of patients through the trial, including recruitment, enrollment, randomization, withdrawal and completion, and a detailed description of the randomization procedure. The CONSORT checklist and template flow diagram are available online.

Registration of clinical trials

Registration in a public trials registry is a condition for publication of clinical trials in this journal in accordance with International Committee of Medical Journal Editors recommendations. Trials must register at or before the onset of patient enrolment. The clinical trial registration number should be included at the end of the abstract of the article. A clinical trial is defined as any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects of health outcomes. Health-related interventions include any intervention used to modify a biomedical or health-related outcome (for example drugs, surgical procedures, devices, behavioural treatments, dietary interventions, and process-of-care changes). Health outcomes include any biomedical or health-related measures obtained in patients or participants, including pharmacokinetic measures and adverse events. Purely observational studies (those in which the assignment of the medical intervention is not at the discretion of the investigator) will not require registration.

Copyright

Upon acceptance of an article, authors will be asked to complete a 'Journal Publishing Agreement' (see more information on this). An e-mail will be sent to the corresponding author confirming receipt of the manuscript together with a 'Journal Publishing Agreement' form or a link to the online version of this agreement.

Subscribers may reproduce tables of contents or prepare lists of articles including abstracts for internal circulation within their institutions. Permission of the Publisher is required for resale or distribution outside the institution and for all other derivative works, including compilations and translations. If excerpts from other copyrighted works are included, the author(s) must obtain written permission from the copyright owners and credit the source(s) in the article. Elsevier has preprinted forms for use by authors in these cases.

For open access articles: Upon acceptance of an article, authors will be asked to complete an 'Exclusive License Agreement' (more information). Permitted third party reuse of open access articles is determined by the author's choice of user license.

Author rights

As an author you (or your employer or institution) have certain rights to reuse your work. More information.

Elsevier supports responsible sharing

Find out how you can share your research published in Elsevier journals.

Role of the funding source

You are requested to identify who provided financial support for the conduct of the research and/or preparation of the article and to briefly describe the role of the sponsor(s), if any, in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication. If the funding source(s) had no such involvement then this should be stated.

Funding body agreements and policies

Elsevier has established a number of agreements with funding bodies which allow authors to comply with their funder's open access policies. Some funding bodies will reimburse the author for the Open Access Publication Fee. Details of existing agreements are available online.

After acceptance, open access papers will be published under a noncommercial license. For authors requiring a commercial CC BY license, you can apply after your manuscript is accepted for publication.

Open access

This journal offers authors a choice in publishing their research:

Open access

- Articles are freely available to both subscribers and the wider public with permitted reuse.
- An open access publication fee is payable by authors or on their behalf, e.g. by their research funder or institution.

Subscription

- Articles are made available to subscribers as well as developing countries and patient groups through our universal access programs.
- No open access publication fee payable by authors.

Regardless of how you choose to publish your article, the journal will apply the same peer review criteria and acceptance standards.

For open access articles, permitted third party (re)use is defined by the following Creative Commons user licenses:

Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

For non-commercial purposes, lets others distribute and copy the article, and to include in a collective work (such as an anthology), as long as they credit the author(s) and provided they do not alter or modify the article.

The open access publication fee for this journal is **USD 3000**, excluding taxes. Learn more about Elsevier's pricing policy: http://www.elsevier.com/openaccesspricing.

Green open access

Authors can share their research in a variety of different ways and Elsevier has a number of green open access options available. We recommend authors see our green open access page for further information. Authors can also self-archive their manuscripts immediately and enable public access from their institution's repository after an embargo period. This is the version that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and in editor-author communications. Embargo period: For subscription articles, an appropriate amount of time is needed for journals to deliver value to subscribing customers before an article becomes freely available to the public. This is the embargo period and it begins from the date the article is formally published online in its final and fully citable form. Find out more.

This journal has an embargo period of 12 months.

Elsevier Publishing Campus

The Elsevier Publishing Campus (www.publishingcampus.com) is an online platform offering free lectures, interactive training and professional advice to support you in publishing your research. The College of Skills training offers modules on how to prepare, write and structure your article and explains how editors will look at your paper when it is submitted for publication. Use these resources, and more, to ensure that your submission will be the best that you can make it.

Language (usage and editing services)
Please write your text in good English (American or British usage is accepted, but not a mixture of these). Authors who feel their English language manuscript may require editing to eliminate possible grammatical or spelling errors and to conform to correct scientific English may wish to use the English Language Editing service available from Elsevier's WebShop.

Informed consent and patient details

Studies on patients or volunteers require ethics committee approval and informed consent, which should be documented in the paper. Appropriate consents, permissions and releases must be obtained where an author wishes to include case details or other personal information or images of patients and any other individuals in an Elsevier publication. Written consents must be retained by the author and copies of the consents or evidence that such consents have been obtained must be provided to Elsevier on request. For more information, please review the Elsevier Policy on the Use of Images or Personal Information of Patients or other Individuals. Unless you have written permission from the patient (or, where applicable, the next of kin), the personal details of any patient included in any part of the article and in any supplementary materials (including all illustrations and videos) must be removed before submission.

Submission

Our online submission system guides you stepwise through the process of entering your article details and uploading your files. The system converts your article files to a single PDF file used in the peer-review process. Editable files (e.g., Word, LaTeX) are required to typeset your article for final publication. All correspondence, including notification of the Editor's decision and requests for revision, is sent by e-mail.

Structure of manuscripts
The following types of contributions will be published: (i) Papers reporting original work; (ii) Interpretative reviews; (iii) Technical notes; (iv) Letters to the Editor

All manuscripts, except Letters to the Editor, should have the following structure:
- Title page, including keywords
- Structured abstract
- Body of the manuscript
- Authors' contributions
- Acknowledgements
- Statement on conflicts of interest
- Summary table
- References
- Appendices (if applicable)
- Maximum word count for research articles: 3000
- Maximum word count for reviews: 4000

Manuscripts not conforming to this structure may be returned to author without prejudice but without review.

Submit your article
Please submit your article via http://ees.elsevier.com/ijmi/.

PREPARATION

NEW SUBMISSIONS

Submission to this journal proceeds totally online and you will be guided stepwise through the creation and uploading of your files. The system automatically converts your files to a single PDF file, which is used in the peer-review process.
As part of the Your Paper Your Way service, you may choose to submit your manuscript as a single file to be used in the refereeing process. This can be a PDF file or a Word document, in any format or lay-out that can be used by referees to evaluate your manuscript. It should contain high enough quality figures for refereeing. If you prefer to do so, you may still provide all or some of the source files at the initial submission. Please note that individual figure files larger than 10 MB must be uploaded separately.

References
There are no strict requirements on reference formatting at submission. References can be in any style or format as long as the style is consistent. Where applicable, author(s) name(s), journal title/book title, chapter title/article title, year of publication, volume number/book chapter and the pagination must be present. Use of DOI is highly encouraged. The reference style used by the journal will be applied to the accepted article by Elsevier at the proof stage. Note that missing data will be highlighted at proof stage for the author to correct.

Formatting requirements
There are no strict formatting requirements but all manuscripts must contain the essential elements needed to convey your manuscript, for example Abstract, Keywords, Introduction, Materials and Methods, Results, Conclusions, Artwork and Tables with Captions.
If your article includes any Videos and/or other Supplementary material, this should be included in your initial submission for peer review purposes.
Divide the article into clearly defined sections.

Please ensure the text of your paper is double-spaced this is an essential peer review requirement.

Figures and tables embedded in text
Please ensure the figures and the tables included in the single file are placed next to the relevant text in the manuscript, rather than at the bottom or the top of the file. The corresponding caption should be placed directly below the figure or table.

Peer review
This journal operates a single blind review process. All contributions will be initially assessed by the editor for suitability for the journal. Papers deemed suitable are then typically sent to a minimum of two independent expert reviewers to assess the scientific quality of the paper. The Editor is responsible for the final decision regarding acceptance or rejection of articles. The Editor's decision is final. More information on types of peer review.

REVISED SUBMISSIONS
Use of word processing software
Regardless of the file format of the original submission, at revision you must provide us with an editable file of the entire article. Keep the layout of the text as simple as possible. Most formatting codes will be removed and replaced on processing the article. The electronic text should be prepared in a way very similar to that of conventional manuscripts (see also the Guide to Publishing with Elsevier). See also the section on Electronic artwork.
To avoid unnecessary errors you are strongly advised to use the 'spell-check' and 'grammar-check' functions of your word processor.

Article structure
Subdivision - numbered sections
Divide your article into clearly defined and numbered sections. Subsections should be numbered 1.1 (then 1.1.1, 1.1.2, ...), 1.2, etc. (the abstract is not included in section numbering). Use this numbering also for internal cross-referencing: do not just refer to 'the text'. Any subsection may be given a brief heading. Each heading should appear on its own separate line.

Introduction
State the objectives of the work and provide an adequate background, avoiding a detailed literature survey or a summary of the results.

Material and methods
Provide sufficient detail to allow the work to be reproduced. Methods already published should be indicated by a reference: only relevant modifications should be described.

Results
Results should be clear and concise.

Discussion
This should explore the significance of the results of the work, not repeat them. A combined Results and Discussion section is often appropriate. Avoid extensive citations and discussion of published literature.

Summary table
The authors shall provide a table with in 2-4 bullets statements on 'what was already known on the topic' and also in 2-4 bullets statements on 'what this study added to our knowledge'. Note that the second part of the table should not list the results of the study as such. It should address what this study has proven and what insights have been gained.

Appendices
If there is more than one appendix, they should be identified as A, B, etc. Formulae and equations in appendices should be given separate numbering: Eq. (A.1), Eq. (A.2), etc.; in a subsequent appendix, Eq. (B.1) and so on. Similarly for tables and figures: Table A.1; Fig. A.1, etc.

Essential title page information

• ***Title.*** Concise and informative. Titles are often used in information-retrieval systems. Avoid abbreviations and formulae where possible.
• ***Author names and affiliations.*** Please clearly indicate the given name(s) and family name(s) of each author and check that all names are accurately spelled. Present the authors' affiliation addresses (where the actual work was done) below the names. Indicate all affiliations with a lower-case superscript letter immediately after the author's name and in front of the appropriate address. Provide the full postal address of each affiliation, including the country name and, if available, the e-mail address of each author.
• ***Corresponding author.*** Clearly indicate who will handle correspondence at all stages of refereeing and publication, also post-publication. **Ensure that the e-mail address is given and that contact details are kept up to date by the corresponding author.**
• ***Present/permanent address.*** If an author has moved since the work described in the article was done, or was visiting at the time, a 'Present address' (or 'Permanent address') may be indicated as a footnote to that author's name. The address at which the author actually did the work must be retained as the main, affiliation address. Superscript Arabic numerals are used for such footnotes.

Structured abstract

A structured abstract, by means of appropriate headings, should provide the context or background for the research and should state its purpose, basic procedures (selection of study subjects or laboratory animals, observational and analytical methods), main findings (giving specific effect sizes and their statistical significance, if possible), and principal conclusions. It should emphasize new and important aspects of the study or observations.

Graphical abstract
Although a graphical abstract is optional, its use is encouraged as it draws more attention to the online article. The graphical abstract should summarize the contents of the article in a concise, pictorial form designed to capture the attention of a wide readership. Graphical abstracts should be submitted as a separate file in the online submission system. Image size: Please provide an image with a minimum of 531 × 1328 pixels (h × w) or proportionally more. The image should be readable at a size of 5 × 13 cm using a regular screen resolution of 96 dpi. Preferred file types: TIFF, EPS, PDF or MS Office files. You can view Example Graphical Abstracts on our information site.
Authors can make use of Elsevier's Illustration and Enhancement service to ensure the best presentation of their images and in accordance with all technical requirements: Illustration Service.

Keywords

Immediately after the abstract, provide a maximum of 6 keywords, using American spelling and avoiding general and plural terms and multiple concepts (avoid, for example, 'and', 'of'). Be sparing with abbreviations: only abbreviations firmly established in the field may be eligible. These keywords will be used for indexing purposes.

Abbreviations
Define abbreviations that are not standard in this field in a footnote to be placed on the first page of the article. Such abbreviations that are unavoidable in the abstract must be defined at their first mention there, as well as in the footnote. Ensure consistency of abbreviations throughout the article.

The abbreviation should be used, except in section headings and subheadings where full text is preferred. Abbreviation should not contain periods or intervening space between letters. Universally known abbreviations (USA for United States of America) need not be defined. Avoid using abbreviations in the abstract of the manuscript.

Acknowledgements
Collate acknowledgements in a separate section at the end of the article before the references and do not, therefore, include them on the title page, as a footnote to the title or otherwise. List here those individuals who provided help during the research (e.g., providing language help, writing assistance or proof reading the article, etc.).

Formatting of funding sources
List funding sources in this standard way to facilitate compliance to funder's requirements:

Funding: This work was supported by the National Institutes of Health [grant numbers xxxx, yyyy]; the Bill & Melinda Gates Foundation, Seattle, WA [grant number zzzz]; and the United States Institutes of Peace [grant number aaaa].

It is not necessary to include detailed descriptions on the program or type of grants and awards. When funding is from a block grant or other resources available to a university, college, or other research institution, submit the name of the institute or organization that provided the funding.

If no funding has been provided for the research, please include the following sentence:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Units
Follow internationally accepted rules and conventions: use the international system of units (SI). If other units are mentioned, please give their equivalent in SI.

Footnotes
Footnotes should be used sparingly. Number them consecutively throughout the article. Many word processors build footnotes into the text, and this feature may be used. Should this not be the case, indicate the position of footnotes in the text and present the footnotes themselves separately at the end of the article.

Artwork

Electronic artwork
General points
- Make sure you use uniform lettering and sizing of your original artwork.
- Preferred fonts: Arial (or Helvetica), Times New Roman (or Times), Symbol, Courier.
- Number the illustrations according to their sequence in the text.
- Use a logical naming convention for your artwork files.
- Indicate per figure if it is a single, 1.5 or 2-column fitting image.
- For Word submissions only, you may still provide figures and their captions, and tables within a single file at the revision stage.
- Please note that individual figure files larger than 10 MB must be provided in separate source files.

A detailed guide on electronic artwork is available.
You are urged to visit this site; some excerpts from the detailed information are given here.
Formats
Regardless of the application used, when your electronic artwork is finalized, please 'save as' or convert the images to one of the following formats (note the resolution requirements for line drawings, halftones, and line/halftone combinations given below):
EPS (or PDF): Vector drawings. Embed the font or save the text as 'graphics'.
TIFF (or JPG): Color or grayscale photographs (halftones): always use a minimum of 300 dpi.
TIFF (or JPG): Bitmapped line drawings: use a minimum of 1000 dpi.
TIFF (or JPG): Combinations bitmapped line/half-tone (color or grayscale): a minimum of 500 dpi is required.
Please do not:
- Supply files that are optimized for screen use (e.g., GIF, BMP, PICT, WPG); the resolution is too low.
- Supply files that are too low in resolution.
- Submit graphics that are disproportionately large for the content.

Color artwork
Please make sure that artwork files are in an acceptable format (TIFF (or JPEG), EPS (or PDF), or MS Office files) and with the correct resolution. If, together with your accepted article, you submit usable color figures then Elsevier will ensure, at no additional charge, that these figures will appear

in color online (e.g., ScienceDirect and other sites) regardless of whether or not these illustrations are reproduced in color in the printed version. **For color reproduction in print, you will receive information regarding the costs from Elsevier after receipt of your accepted article**. Please indicate your preference for color: in print or online only. Further information on the preparation of electronic artwork.

Illustration services
Elsevier's WebShop offers Illustration Services to authors preparing to submit a manuscript but concerned about the quality of the images accompanying their article. Elsevier's expert illustrators can produce scientific, technical and medical-style images, as well as a full range of charts, tables and graphs. Image 'polishing' is also available, where our illustrators take your image(s) and improve them to a professional standard. Please visit the website to find out more.

Figure captions
Ensure that each illustration has a caption. A caption should comprise a brief title (**not** on the figure itself) and a description of the illustration. Keep text in the illustrations themselves to a minimum but explain all symbols and abbreviations used.

Tables

Please submit tables as editable text and not as images. Tables can be placed either next to the relevant text in the article, or on separate page(s) at the end. Number tables consecutively in accordance with their appearance in the text and place any table notes below the table body. Be sparing in the use of tables and ensure that the data presented in them do not duplicate results described elsewhere in the article. Please avoid using vertical rules and shading in table cells.

References

Citation in text
Please ensure that every reference cited in the text is also present in the reference list (and vice versa). Any references cited in the abstract must be given in full. Unpublished results and personal communications are not recommended in the reference list, but may be mentioned in the text. If these references are included in the reference list they should follow the standard reference style of the journal and should include a substitution of the publication date with either 'Unpublished results' or 'Personal communication'. Citation of a reference as 'in press' implies that the item has been accepted for publication.

Reference links
Increased discoverability of research and high quality peer review are ensured by online links to the sources cited. In order to allow us to create links to abstracting and indexing services, such as Scopus, CrossRef and PubMed, please ensure that data provided in the references are correct. Please note that incorrect surnames, journal/book titles, publication year and pagination may prevent link creation. When copying references, please be careful as they may already contain errors. Use of the DOI is encouraged.

A DOI can be used to cite and link to electronic articles where an article is in-press and full citation details are not yet known, but the article is available online. A DOI is guaranteed never to change, so you can use it as a permanent link to any electronic article. An example of a citation using DOI for an article not yet in an issue is: VanDecar J.C., Russo R.M., James D.E., Ambeh W.B., Franke M. (2003). Aseismic continuation of the Lesser Antilles slab beneath northeastern Venezuela. Journal of Geophysical Research, https://doi.org/10.1029/2001JB000684. Please note the format of such citations should be in the same style as all other references in the paper.

Web references
As a minimum, the full URL should be given and the date when the reference was last accessed. Any further information, if known (DOI, author names, dates, reference to a source publication, etc.), should also be given. Web references can be listed separately (e.g., after the reference list) under a different heading if desired, or can be included in the reference list.

Data references
This journal encourages you to cite underlying or relevant datasets in your manuscript by citing them in your text and including a data reference in your Reference List. Data references should include the following elements: author name(s), dataset title, data repository, version (where available), year, and global persistent identifier. Add [dataset] immediately before the reference so we can properly identify it as a data reference. The [dataset] identifier will not appear in your published article.

References in a special issue
Please ensure that the words 'this issue' are added to any references in the list (and any citations in the text) to other articles in the same Special Issue.

Reference management software
Most Elsevier journals have their reference template available in many of the most popular reference management software products. These include all products that support Citation Style Language styles, such as Mendeley and Zotero, as well as EndNote. Using the word processor plug-ins from these products, authors only need to select the appropriate journal template when preparing their article, after which citations and bibliographies will be automatically formatted in the journal's style. If no template is yet available for this journal, please follow the format of the sample references and citations as shown in this Guide.

Users of Mendeley Desktop can easily install the reference style for this journal by clicking the following link:
http://open.mendeley.com/use-citation-style/international-journal-of-medical-informatics
When preparing your manuscript, you will then be able to select this style using the Mendeley plug-ins for Microsoft Word or LibreOffice.

Reference formatting
There are no strict requirements on reference formatting at submission. References can be in any style or format as long as the style is consistent. Where applicable, author(s) name(s), journal title/book title, chapter title/article title, year of publication, volume number/book chapter and the pagination must be present. Use of DOI is highly encouraged. The reference style used by the journal will be applied to the accepted article by Elsevier at the proof stage. Note that missing data will be highlighted at proof stage for the author to correct. If you do wish to format the references yourself they should be arranged according to the following examples:

Reference style
Text: Indicate references by number(s) in square brackets in line with the text. The actual authors can be referred to, but the reference number(s) must always be given.
Example: '..... as demonstrated [3,6]. Barnaby and Jones [8] obtained a different result ....'
List: Number the references (numbers in square brackets) in the list in the order in which they appear in the text.
Examples:
Reference to a journal publication:
[1] J. van der Geer, J.A.J. Hanraads, R.A. Lupton, The art of writing a scientific article, J. Sci. Commun. 163 (2010) 51–59.
Reference to a book:
[2] W. Strunk Jr., E.B. White, The Elements of Style, fourth ed., Longman, New York, 2000.
Reference to a chapter in an edited book:
[3] G.R. Mettam, L.B. Adams, How to prepare an electronic version of your article, in: B.S. Jones, R.Z. Smith (Eds.), Introduction to the Electronic Age, E-Publishing Inc., New York, 2009, pp. 281–304.
Reference to a website:
[4] Cancer Research UK, Cancer statistics reports for the UK. http://www.cancerresearchuk.org/aboutcancer/statistics/cancerstatsreport/, 2003 (accessed 13.03.03).
Reference to a dataset:
[dataset] [5] M. Oguro, S. Imahiro, S. Saito, T. Nakashizuka, Mortality data for Japanese oak wilt disease and surrounding forest compositions, Mendeley Data, v1, 2015. https://doi.org/10.17632/xwj98nb39r.1.

Journal abbreviations source
Journal names should be abbreviated according to the List of Title Word Abbreviations.

The referencing style as used by the NLM are preferred (see http://www.nlm.nih.gov/bsd/uniform_requirements.html for examples) Abbreviations for journals are those used in MeSH published by the US National Library of Medicine http://www.ncbi.nlm.nih.gov/nlmcatalog/journals.

Video

Elsevier accepts video material and animation sequences to support and enhance your scientific research. Authors who have video or animation files that they wish to submit with their article are strongly encouraged to include links to these within the body of the article. This can be done in the same way as a figure or table by referring to the video or animation content and noting in the body text where it should be placed. All submitted files should be properly labeled so that they directly

relate to the video file's content. In order to ensure that your video or animation material is directly usable, please provide the files in one of our recommended file formats with a preferred maximum size of 150 MB. Video and animation files supplied will be published online in the electronic version of your article in Elsevier Web products, including ScienceDirect. Please supply 'stills' with your files: you can choose any frame from the video or animation or make a separate image. These will be used instead of standard icons and will personalize the link to your video data. For more detailed instructions please visit our video instruction pages. Note: since video and animation cannot be embedded in the print version of the journal, please provide text for both the electronic and the print version for the portions of the article that refer to this content.

Supplementary material

Supplementary material such as applications, images and sound clips, can be published with your article to enhance it. Submitted supplementary items are published exactly as they are received (Excel or PowerPoint files will appear as such online). Please submit your material together with the article and supply a concise, descriptive caption for each supplementary file. If you wish to make changes to supplementary material during any stage of the process, please make sure to provide an updated file. Do not annotate any corrections on a previous version. Please switch off the 'Track Changes' option in Microsoft Office files as these will appear in the published version.

AudioSlides

The journal encourages authors to create an AudioSlides presentation with their published article. AudioSlides are brief, webinar-style presentations that are shown next to the online article on ScienceDirect. This gives authors the opportunity to summarize their research in their own words and to help readers understand what the paper is about. More information and examples are available. Authors of this journal will automatically receive an invitation e-mail to create an AudioSlides presentation after acceptance of their paper.

AFTER ACCEPTANCE

Accepted Manuscripts

As a service to the community, this journal makes available online the accepted manuscripts as soon as possible after acceptance. At this stage, the author's accepted manuscript (in both full-text and PDF) is given a Digital Object Identifier (DOI) and is fully citable, and searchable by title, author(s) name and the full-text. The article also carries a disclaimer noting that it is an unedited manuscript which has not yet been copyedited, typeset or proofread. When the fully copyedited version is ready for publication, it simply replaces the author accepted manuscript version.

Online proof correction

Corresponding authors will receive an e-mail with a link to our online proofing system, allowing annotation and correction of proofs online. The environment is similar to MS Word: in addition to editing text, you can also comment on figures/tables and answer questions from the Copy Editor. Web-based proofing provides a faster and less error-prone process by allowing you to directly type your corrections, eliminating the potential introduction of errors.

If preferred, you can still choose to annotate and upload your edits on the PDF version. All instructions for proofing will be given in the e-mail we send to authors, including alternative methods to the online version and PDF.

We will do everything possible to get your article published quickly and accurately. Please use this proof only for checking the typesetting, editing, completeness and correctness of the text, tables and figures. Significant changes to the article as accepted for publication will only be considered at this stage with permission from the Editor. It is important to ensure that all corrections are sent back to us in one communication. Please check carefully before replying, as inclusion of any subsequent corrections cannot be guaranteed. Proofreading is solely your responsibility.

Book Reviews

Publishing houses interested in having their books reviewed should contact our editor for reviews Patrice Degoulet (patrice.degoulet@egp.aphp.fr).

Offprints

The corresponding author will, at no cost, receive a customized Share Link providing 50 days free access to the final published version of the article on ScienceDirect. The Share Link can be used for sharing the article via any communication channel, including email and social media. For an extra charge, paper offprints can be ordered via the offprint order form which is sent once the article is accepted for publication. Both corresponding and co-authors may order offprints at any

time via Elsevier's Webshop. Corresponding authors who have published their article open access do not receive a Share Link as their final published version of the article is available open access on ScienceDirect and can be shared through the article DOI link.

AUTHOR INQUIRIES

Visit the Elsevier Support Center to find the answers you need. Here you will find everything from Frequently Asked Questions to ways to get in touch.
You can also check the status of your submitted article or find out when your accepted article will be published.

© Copyright 2014 Elsevier | http://www.elsevier.com

ANEXO C – Artigo publicado em revista internacional

International Journal of Computer and Information Technology (ISSN: 2279 – 0764)
Volume 05 – Issue 05, September 2016

Text Mining as a Strategy in Profiling the Use of Influenza Virus Genome in Scientific Publications

Fernanda C. R. Correa
Federal University of Health Sciences of Porto Alegre (UFCSPA)
Porto Alegre, Brazil
Email: fernandacr [AT] ufcspa.edu.br

Aline A. Vanin
Federal University of Health Sciences of Porto Alegre (UFCSPA)
Porto Alegre, Brazil

Silvio C. Cazella
Federal University of Health Sciences of Porto Alegre (UFCSPA)
Porto Alegre, Brazil

Abstract — **The aim of this study was to profile the use and usage patterns of influenza virus genome from scientific publications in online databases using Natural Language Processing and Text Mining techniques. A systematic research was performed to select papers in PubMed electronic database using the keywords: 'influenza', 'genome', 'database'. The 45 articles that presented free full text available were processed with the sofwares AntFileConverter and AntConc. Text Mining was performed with the software Weka. Association rules were expected between genome and influenza. Also, it was predicted that influenza genome and terms related directly to the application of genome databases would relate. However, the results revealed an association between influenza virus protein and mutation sequence/database. The discovery of different associations than the expected revealed the necessity of expanding the research in order to increase the size of the corpus and to improve the attributes selection for mining in Weka sofware.**

Keywords – Data Mining; Natural Language Processing; Influenza A virus; Genome, Viral; Databases, Nucleic Acid

I. INTRODUCTION

New approaches based upon molecular and computational methods are essential for advances in the study and control of infectious diseases. In this context, pathogenic viruses such as influenza viruses are an important source of study for the development of these new methods, given the large amount of information available on these microorganisms [1].

Influenza virus is an important human pathogen that causes a high number of deaths every year. Seasonal influenza epidemics result in over three million cases of severe illness, and about 250,000 to 500,000 deaths every year [2]. Influenza virus has an annual attack rate estimated at 5-10% in adults and 20-30% in children, with possible hospitalization and death specially in more susceptible population such as the elderly, the chronically ill, and pregnant women [3].

Prevention and control of influenza epidemics is a major problem for public health care services [4]. The current approach is annual vaccination with a trivalent inactivated influenza vaccine, composed by two different influenza type A strains and one influenza type B strain [5].

Influenza A virus has a high mutation rate and wide host range. Avian H5N1 and H7N9 were able to cause human infections, raising the fear of a new influenza strain that could result in a global pandemic due to the absence of previous host immunity [6] [7]. However, due to constant epidemiological control by health authorities, there is more control of a possible outbreak. [8] [9] [10].

Biological data from online databases provide an excellent source of material for research. Genomes sequences are easily and quickly obtained [11]. Due to the recent advent of methodologies allowing the analysis of these materials, mutations can be identified and used to predict the emergence of new strains with pathogenic potential as well as to understand how viruses spread geographically [12].

The data generated and accumulated in biological databases is consistent and abundant, creating an overload of information. Thus, computational techniques and new technologies are necessary to provide effective and efficient analysis of this content. Moreover, it is important to evaluate the amount of information that is generated and also to detect the potential of knowledge discovery that result from the application of these technologies [1].

Natural Language Processing (NLP) and Text Mining are two instruments that can be used in order to identify and extract relevant information from medical journals. The NLP consists on processing natural language texts by computer to access their meaning. Text Mining is a variation on a field called Data Mining which discovers and extracts knowledge from data, comprising activities such as information retrieval, information extraction and data mining to find associations among the pieces of information extracted from many different texts [13].

Text Mining, also known as Knowledge-Discovery in Texts (KDT), refers generally to the extraction of interesting and non-trivial information and knowledge from unstructured text. KDT combines extraction techniques, information retrieval, NLP and summarization of documents with data mining methods. Text Mining systems have been applied to the biological research area since the late 1990s and have considerably improved since then [14] [15].

International Journal of Computer and Information Technology (ISSN: 2279 – 0764)
Volume 05 – Issue 05, September 2016

Association rule mining is a technique used to discover relationships among a large set of variables in a data set. In association rules for text mining, the focus is to study relationships and implications among topics that are used to characterize a corpus, aiming to discover relevant association rules within a corpus such as the presence of a set of terms in an article implying the presence of another term [14].

The aim of this study was to analyze scientific publications available online to verify the use of influenza virus genome from online databases using NLP and Text Mining techniques. Moreover, associations or usage patterns of influenza genomes for different purposes are expected to be discovered. The results will provide useful information regarding the scientific publications generated from the study of influenza virus genome and thus displaying the potential use of these data in future studies and publications.

II. METHODS

For this systematic analysis, the PubMed electronic database was researched by using the following keywords: 'influenza', 'genome', 'database'. A total of 76 results were obtained, and 45 articles that presented free full text available (using PubMed filter) were selected. No other keywords or restrictions were used.

The .pdf files were converted into .txt files using the AntFileConverter 1.0.0 software [16]. The text files were cleaned, for removal of abstracts, titles, author informations and references. Next, a list of the 30 most common and relevant words was selected with the corpus analysis software AntConc 3.4.3 [17]. The texts were tokenized using binary representation in a Microsoft Office Excel 2010 spreadsheet. An Attribute-Relation File Format (ARFF) file with 30 attributes and 45 data instances was created. The presence of attributes was indicated as "1" (number ONE), while null values were indicated as "?" (question mark). A partial representation of the ARFF file is shown in Figure 1.

Text mining was performed with the software Weka 3.6 [18] in order to find association rules using the Apriori algorithm at minimum support of 0.6 and lift of 1.1. The software generated ten rules, as shown in Figure 2.

III. RESULTS AND DISCUSSION

The support of an itemset is defined as the proportion of transactions in the data set which contain the itemset. In this database, a set of items only appeared as a rule if it has occurred in 60% (0,6) of all transactions. The lift metrics is able to assess whether two items are positively or negatively independent, and also determines if two items are independent of each other. A minimum lif of 1.1 will only select rules in which the items are positively independent of each other [19].

The best rules showing the most frequently associated words are the rules 2, 5, 6 and 8, presenting higher conviction values than the others. Lift values are the same. Conviction, confidence and leverage values are very similar among this four rules. The elevated values of conviction indicate a higher independence of the foregoing item in relation to the consequent item. The rules are resemblant, indicating an association between influenza virus protein and mutation sequence/database. The most interesting rules are shown in Table 1.

```
@relation terms
@attribute Sequence {?,1}
@attribute Influenza {?,1}
@attribute Virus {?,1}
@attribute Gene {?,1}
@attribute Protein {?,1}
@attribute Database {?,1}
@attribute Genome {?,1}
@attribute Genbank {?,1}
@attribute Sequencing {?,1}
@attribute Pandemic {?,1}
@attribute Mutation {?,1}
@attribute Genotype {?,1}
@attribute Research {?,1}
@attribute Antigenic {?,1}
@attribute Pcr {?,1}
@attribute Vaccine {?,1}
@attribute Method {?,1}
@attribute Drug {?,1}
@attribute Antiviral {?,1}
@attribute Resistance {?,1}
@attribute Synthesis {?,1}
@attribute Assay {?,1}
@attribute Primer {?,1}
@attribute Openflu {?,1}
@attribute Array {?,1}
@attribute Culture {?,1}
@attribute Transcriptome {?,1}
@attribute Proteome {?,1}
@attribute Diagnostic {?,1}
@attribute Phylogenetic {?,1}

@data
%
% 45 instances
%
1,1,1,?,1,1,1,1,1,1,1,1,?,?,?,?,1,?,1,?,?,?,?,1,1,?,?,?
1,1,1,1,?,1,1,1,1,1,?,1,1,1,?,?,1,1,?,?,?,?,?,?,?,?,?,1,1
1,1,1,1,?,1,1,?,1,?,1,?,1,?,?,1,1,?,?,?,?,1,?,?,?,?,?,?,?
1,1,1,1,1,1,1,?,?,1,?,?,1,1,?,?,1,?,?,?,?,?,?,?,?,?,?,?,1
1,1,1,1,1,1,1,1,1,1,1,?,1,1,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
```

Figure 1. Partial ARRF File

Considering the amount of selected terms and the content of the papers, it was expected that association rules were

International Journal of Computer and Information Technology (ISSN: 2279 – 0764)
Volume 05 – Issue 05, September 2016

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 1 -C 1.1 -D 0.05 -U 1.0 -M 0.6 -S -1.0 -c -1
Relation:     terms
Instances:    45
Attributes:   30
              Sequence, Influenza, Virus, Gene, Protein, Database, Genome, Genbank, Sequencing, Pandemic, Mutation, Genotype,
              Research, Antigenic, Pcr, Vaccine, Method, Drug, Antiviral, Resistance, Synthesis, Assay, Primer, Openflu, Array, Culture,
              Transcriptome, Proteome, Diagnostic, Phylogenetic

=== Associator model (full training set) ===

Apriori
=======

Minimum support: 0.6 (27 instances)
Minimum metric <lift>: 1.1
Number of cycles performed: 8

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9
Size of set of large itemsets L(2): 32
Size of set of large itemsets L(3): 55
Size of set of large itemsets L(4): 51
Size of set of large itemsets L(5): 27
Size of set of large itemsets L(6): 8
Size of set of large itemsets L(7): 1

Best rules found:

 1. Influenza=1 Virus=1 Protein=1 38 ==> Mutation=1 28    conf:(0.74) < lift:(1.14)> lev:(0.08) [3] conv:(1.23)
 2. Mutation=1 29 ==> Influenza=1 Virus=1 Protein=1 28    conf:(0.97) < lift:(1.14)> lev:(0.08) [3] conv:(2.26)
 3. Influenza=1 Virus=1 Protein=1 38 ==> Database=1 Mutation=1 28    conf:(0.74) < lift:(1.14)> lev:(0.08) [3] conv:(1.23)
 4. Influenza=1 Virus=1 Protein=1 Database=1 38 ==> Mutation=1 28    conf:(0.74) < lift:(1.14)> lev:(0.08) [3] conv:(1.23)
 5. Mutation=1 29 ==> Influenza=1 Virus=1 Protein=1 Database=1 28    conf:(0.97) < lift:(1.14)> lev:(0.08) [3] conv:(2.26)
 6. Database=1 Mutation=1 29 ==> Influenza=1 Virus=1 Protein=1 28    conf:(0.97) < lift:(1.14)> lev:(0.08) [3] conv:(2.26)
 7. Influenza=1 Virus=1 Protein=1 38 ==> Sequence=1 Mutation=1 27    conf:(0.71) < lift:(1.14)> lev:(0.07) [3] conv:(1.2)
 8. Sequence=1 Mutation=1 28 ==> Influenza=1 Virus=1 Protein=1 27    conf:(0.96) < lift:(1.14)> lev:(0.07) [3] conv:(2.18)
 9. Influenza=1 Virus=1 Protein=1 38 ==> Sequence=1 Database=1 Mutation=1 27    conf:(0.71) < lift:(1.14)> lev:(0.07) [3]
    conv:(1.2)
10. Influenza=1 Virus=1 Protein=1 Database=1 38 ==> Sequence=1 Mutation=1 27    conf:(0.71) < lift:(1.14)> lev:(0.07) [3]
    conv:(1.2)
```

Figure 2. Data mining results

discovered between genome and influenza. The term genome was one of the research words and was not present in any of the association rules, even though it was present on 42 of the 45 papers. Also, it was predicted that association between influenza genome and terms related directly to the application of genome databases in research of vaccine development, drug testing and development, assays and production of identification and diagnostics tests, assays for phylogenetic identification, and studies in genomics, proteomics and transcriptomics.

International Journal of Computer and Information Technology (ISSN: 2279 – 0764)
Volume 05 – Issue 05, September 2016

TABLE I. MOST INTERESTING ASSOCIATION RULES

Rules	*Confidence*	*Lift*	*Leverage*	*Conviction*
IF Mutation THEN Influenza AND Virus AND Protein	0.97	1.14	0.08	2.26
IF Mutation THEN Influenza AND Virus AND Protein AND Database	0.97	1.14	0.08	2.26
IF Database AND Mutation THEN Influenza AND Virus AND Protein	0.97	1.14	0.08	2.26
IF Sequence AND Mutation THEN Influenza AND Virus AND Protein	0.96	1.14	0.07	2.18

These unexpected results could be a consequence of the size of the corpus. However, it is not known if a corpus with all the publications from the initial results (n=76) would produce different association rules than the ones generated from the current corpus (n=45).

As in regard of the PubMed systematic research, it is possible to verify a low scientific production on the area. Previous to the search of the terms 'influenza' 'genome' 'database', a search using Medical Subject Heading (MeSH) terms was performed with fewer results, as shown in Table 2. Even with the use of generic terms instead of MeSH terms, the amount of results obtained is still considered low. This demonstrates that influenza genome data is not being used to its full potential. Furthermore, we can also question the methods that the journals and the authors are using to index their publications. Articles with an inappropriate selection of MeSH terms will result in a possible deficient recuperation of this material. The amount of information available from online influenza genome databases would provide data for research for a range of studies, from the discovery of new treatment drugs and vaccines to the widening of epidemiological studies.

TABLE II. PUBMED MESH TERMS RESEARCH RESULTS

MeSH Terms	*Number of results*
Influenza A virus AND Genome, Viral AND Databases, Genetic	26
Influenza A virus AND Genome, Viral AND Databases, Nucleic Acid	7
Influenza A virus AND Genome, Viral AND Databases as Topic	28
Influenza A virus AND Genome AND Databases, Genetic	30
Influenza A virus AND Genome AND Databases, Nucleic Acid	7
Influenza A virus AND Genome AND Databases as Topic	35
Influenza, Human AND Genome, Viral AND Databases, Genetic	8
Influenza, Human AND Genome, Viral AND Databases, Nucleic Acid	3
Influenza, Human AND Genome, Viral AND Databases as Topic	8
Influenza, Human AND Genome AND Databases, Genetic	9
Influenza, Human AND Genome AND Databases, Nucleic Acid	3
Influenza, Human AND Genome AND Databases as Topic	11

IV. CONCLUSION

In order to increase the frequency of terms and to generate different rules associating influenza genome databases and its applications, medical ontologies such as the Gene Ontology [20] and the Unified Medical Language System (UMLS) [21] can be used. Also, the use of the regular expressions (regex) system from the AntConc software can increase the efficiency of the selection of terms. Moreover, a systematic research in different electronic databases than PubMed can also increase the corpus.

The results revealed different associations than the expected and several hypotheses emerged. Mining the texts with Weka software highlighted the association between influenza virus protein and mutation sequence/database. This could be a consequence of the corpus size which is an aftermath of the insufficient amount of publications in the field.

To verify these questions, more study is necessary. More data can be obtained by expanding the search to different electronic databases and better results will be generated with the use of medical ontologies and regex. The research to assess the usage profile of influenza genome databases is important to determine its potential applications, such as studies about possible new mutated strains, and researches to develop more efficient vaccines.

REFERENCES

[1] Yang X, Yang H, Zhou G, Zhao, G. Infectious Disease in the Genomic Era. Annu Rev Genomics Hum Genet 2008;9:21-48.

[2] World Health Organization. Influenza (seasonal) fact sheet n°211. 2014.

[3] Pastore APW, Prates C, Gutierrez LLP. Implications of influenza A/H1N1 in gestational period. Implicações da influenza A/H1N1 no período gestacional. Sci Med 2012;22(1):53-8.

[4] Keitel WA, Cate TR, Couch RB. Efficacy of sequential annual vaccination with inactivated influenza virus vaccine. Am J Epidemiol 1988;127:353-64.

[5] Tripp RA, Tompkins SM. Recombinant vaccines for influenza virus. Curr Opin Investig Drugs 2008;9(8):836-45.

[6] Gao R, Cao B, Hu Y, Feng Z, Wang D, et al. Human infection with a novel avian-origin influenza A (H7N9) virus. New England J Med 2013;368:1888-97.

[7] Zhou J, Wang D, Gao R, Zhao B, Song J, et al. Biological features of novel avian influenza A (H7N9) virus. Nature 2013;499:500-3.

[8] Steinhauer DA. Influenza: pathways to human adaptation. Nature 2013;499:412-3.

[9] Ebrahimi M, Aghagolzadeh P, Shamsbadi N, Tahmasebi A, Alsharifi M, et al. Understanding the underlaying mechanism of HA-subtyping in the level of physic-chemical characteristics of protein. PLoS One 2014;9(5):e96984.

[10] Jaskulski PR, Jaskulski MR, Guilhermano LG. Comparison between 1918 and 2009 flu pandemics in São Vicente do Paulo Hospital. Comparação entre as pandemias de gripe de 1918 e 2009 na perspectiva do Hospital São Vicente de Paulo em Passo Fundo, Rio Grande do Sul. Sci Med 2012;22(3):169-74.

[11] He C, Han G, Wang D, Liu W, Li G, Liu X, Ding N. Homologous recombination evidence in human and swine influenza A viruses. Virology 2008;380:12-20.

[12] Nelson MI, Simonsen L, Viboud C, Miller MA, Holmes EC. Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. PLoS Pathog 2007;3:1220-8.

[13] Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends Biotechnol 2006;24(12): 571-9.

[14] Gupta V, Lehal GS. A survey of text mining techniques and applications. J Emerg Technol Web Intell 2009;1(1):60-76.

[15] Erhardt RAA, Schneider R, Blaschke C. Status of text-mining techniques applied to biomedical text. Drug Discov Today 2006;11(7-8):315-25.

[16] Anthony L. AntFileConverter (Version 1.0.0) [Computer Software]. Tokyo, Japan: Waseda University. 2013.

[17] Anthony, L. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. 2014.

[18] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009;11(1):10-8.

[19] Gonçalves EC. Data mining with WEKA. Data mining com a ferramenta WEKA. III Fórum de Software Livre de Duque de Caxias. 2011.

[20] Smith B, Williams J, Schulze-Kremer S. The ontology of the Gene Ontology. Proceeding of the AMIA Symposium 2003;609-13.

[21] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281-91.

ANEXO D – Resumo publicado em anais de evento

IV Escola Regional de Computação Aplicada à Saúde
7 e 8 de Outubro de 2016

Mineração de Dados como ferramenta para descoberta e validação de informações epidemiológicas sobre o vírus influenza A

Fernanda C. R. Corrêa[1], Ana B. G. da Veiga[2], Sílvio C. Cazella[3]

[1]Programa de Pós-Graduação em Ciências da Saúde – Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA)
Rua Sarmento Leite, 245 – 90.050-170 – Porto Alegre – RS – Brazil

[2]Departamento de Ciências Básicas da Saúde – Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA)

[3]Departamento de Ciências Exatas e Sociais Aplicadas – Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA)

{fernandacr,anabgv,silvioc}@ufcspa.edu.br

Resumo. *Técnicas de mineração de dados e os softwares disponíveis permitem a exploração de grandes quantidades de dados à procura de padrões consistentes. Este estudo teve por objetivo analisar um banco de dados com informações sobre cepas isoladas e sequenciadas do vírus influenza A, obtido do site GenBank [National Library of Medicine 2016], procurando identificar padrões que pudessem identificar e/ou validar informações epidemiológicas utilizando ferramentas de mineração de dados. Após processamento, foram selecionados dados de cepas isoladas no período de 2005-2015, resultando em um dataset com 157.639 instâncias, que foi analisado com o software Weka [Witten and Frank 2000]. Após tarefa de associação com o algoritmo Apriori, os resultados demonstraram associação entre o vírus influenza subtipo H1N1 e seres humanos em 2009, com percentual de ocorrência da associação e confiabilidade de 89%. Além disso, também evidenciaram forte associação entre o influenza A H1N1 pandêmico e seres humanos em 2009, com percentual de ocorrência da associação e confiabilidade de 96%, mostrando que o vírus influenza tipo A subtipo H1N1 prevalente em 2009 tinha alta afinidade pelo hospedeiro humano, mesmo sendo este subtipo capaz de infectar outras espécies, como aves, porcos e outros animais [CDC 2014]. A pandemia de influenza A/H1N1 (2009) resultou da emergência de um vírus influenza novo, para o qual a maioria das pessoas não tinha imunidade prévia [WHO 2010], e se espalhou rapidamente pelo mundo, causando a morte de mais de 18 mil pessoas em mais de 200 países [CDC 2009]. A ferramenta testada foi eficiente para analisar um banco de dados público e gerar regras com grande percentual de associabilidade e confiabilidade, com resultados compatíveis com dados epidemiológicos oficiais [WHO 2010]. O estudo realizado evidenciou o potencial da Mineração de Dados em identificar e validar dados epidemiológicos, sendo uma ferramenta promissora para a descoberta de novas informações sobre agentes patogênicos.*

References

CDC – Center for Disease Control and Prevention. (2009) Swine Influenza A (H1N1) Infection in Two Children - Southern California, March-April 2009, April 24. MMWR 58:400-402.

CDC – Center for Disease Control and Prevention. (2014) Types of Influenza Viruses. Disponivel em: <http://www.cdc.gov/flu/about/viruses/types.htm>.

U.S. National Library of Medicine. (2016) FTP access to GenBank data. Disponivel em: <http://www.ncbi.nlm.nih.gov/genbank/ftp>.

WHO – World Health Organization. (2010) Pandemic (H1N1) 2009 - update 107. Global Alert and Response (GAR)/Disease Outbreak News/Weekly Update, 2 Jul.

Witten, Ian H.; Frank, Eibe. (2000) Data mining: practical machine learning tools and techniques with java implementations, San Francisco: Morgan Kaufmann.