Luís Filipe Tavares da Silva

Bioinformática e Aplicações em Virologia

Universidade Fernando Pessoa,

Faculdade Ciências da Saúde.

Porto 2015

Luís Filipe Tavares da Silva

Bioinformática e Aplicações em Virologia

Universidade Fernando Pessoa,

Faculdade Ciências da Saúde.

Porto 2015

Bioinformática e Aplicações em Virologia

Autor: Luís Filipe Tavares da Silva

Orientador: Prof. Doutor Ricardo Magalhães

Trabalho apresentado à Universidade Fernando Pessoa
como parte dos requisitos para obtenção do grau de
Mestre em Ciências Farmacêuticas.

RESUMO

Na década de 1940 foi inventado o primeiro computador. A informação construía-se ligando e desligando o sistema binário. Na década de 50, a descoberta da dupla hélice no ADNveio provar que a informação genética também é escrita ligando e desligando, não um sistema binário como o computador digital, mas quaternário. Mas, só foi nos anos 90 que a informática e a biologia juntaram-se, inicialmente para decifrar a estrutura do ADN.

A virologia é uma das áreas de relevância na bioinformática pelo facto dos vírus serem uma das estruturas mais elementares da vida e ao mesmo tempo, complexas. Com o intuito de facilitar o estudo e a compreensão da diversidade de vírus, foram desenvolvidas muitas classificações mas, na actualidade utiliza-se sobretudo a taxonomia do ICTV e, a combinação entre a classificação de Baltimore e LHT. O estudo dos vírus tem permitido conhecer a forma como estes utilizam os recursos energéticos das células hospedeiros e, no esclarecimento de muitos dos mecanismos biológicos dos próprios hospedeiros. Isto tem sido possível pelos desenvolvimentos tecnológicos e científicos muito significativos nas áreas da biologia molecular e da genética, nomeadamente, nas estruturas do ADN, ARN, do genoma e da sua sequenciação.

A sequenciação é uma série de processos bioquímicos para a determinação da ordem dos nucleótidos. Com o avanço da tecnologia, o número de genomas sequenciados e as bases de dados necessárias para o seu estudotem vindo a aumentar rapidamente, resultando no desenvolvimento exponencial da bioinformática e promovendo a criação de centros de armazenamento e processamento de dados.

A análise bioinformática nos vírus compreende tarefas relacionadas com a análise de sequências novas, incluindo identificação, anotação funcional e análise das relações filogenéticas e, para o seu estudo, existem uma grande diversidade e complexidade de ferramentas que são agrupadas em: identificação ORF e previsão de genes; procura de homogeneidade e alinhamento de sequências; reconhecimento de padrões de epítopos; análise de repetições em tandem curto; domínio transmembranar; estudos estruturais secundários e terciários; análise das vias metabólicas; e analise de dados de microarrays.

Os novos desenvolvimentos, sejam bases de dados e/ou ferramentas, estão a crescer em função do aumento da procura. No entanto, ainda existem limitações, conforme foi possível validar ao longo da leitura científica sobre este tema.

O objetivo deste trabalho foi levar a cabo a identificaçãodas aplicações desenvolvidas dentro da bioinformática para conhecer melhor os vírus.Espera-se que esta visão resumida, - mas muito complexadesde o olhar do farmacêutico- tenha conseguido revelar a importância e utilidade dos recursos bioinformáticos disponíveis para a investigação em virologia.

ABSTRACT

The first computer was invented in the 40s. The information was built connecting and disconnecting a binary system. In the 50s, the discovery of the double helix of the DNA proved that the genetic information was also written connecting and disconnecting but a quaternary system. However, only in the 90s the informatics and the biology joined motivated for the need of decipheringthe structure of DNA.

The Virology is one of the most relevant areas of study of bioinformatics because even the viruses are the most elementary alive structures, at the same time they are complex. Because of the diversity of virus, several classifications were developed, but currently, the ICTV taxonomy, the Baltimore classification and LHT are being used. The bioinformatics provide tools to understand how viruses profit the energy of the host cells and the biological mechanisms of these hosts, mainly through the technological and scientific developments of molecular biology and genetics, specifically the structures of DNA and RNA, genome and sequencing.

The sequencing is a series of biochemical processes to determine the order of nucleoids. The sequenced genomes and the virus data bases are rapidly increasing. As consequence, the bioinformatics shows exponential advances and the store centers and data processing are in the increasing.

Main bioinformatics tasks for virus comprehension are: identification, functional annotation and analysis of phylogenetic relations. It is possible because of the development of tools grouped for: open reading frame identification and gene prediction; homology searching and sequence alignment; patter/motif/epitope recognition; short tandem repeats; transmembrane domains; secondary and tertiary structural studies; pathway analysis; and microarray data analysis.

The data bases and the tools are increasing as response to the growing demand. However, there are still problems to be solved, as verified in the scientific literature.

The objective of this study was the identification of the bioinformatics applications for the study of viruses. This summarized review ï even highly complex from the look of the pharmaceutic - was intended to reveal the usefulness of bioinformatics resources available for virology research.

AGRADECIMENTOS

Agradeço à Universidade Fernando Pessoa, e em especial ao Prof. Doutor Ricardo Magalhães, toda a disponibilidade prestada, atenção dedicada, e apoio, na realização e revisão do presente trabalho. À minha família, o apoio incondicional e a oportunidade que me proporcionaram, ao longo do meu percurso académico, que despoletaramo maior proveito de toda a aprendizagem obtida na Universidade Fernando Pessoa.

Aos meus amigos que foram um grande apoio e me incentivaram sempre.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

ABREVIATURAS

ADN ï ácido desoxirribonucleico

A ï Adenina

Bp ï pares de bases

C ï Citosina

ADN ï ÁcidoDesoxirribonucleico

ExPASy ï Expert Protein Analyses System

EMBL- Laboratório europeu de biologia molecular

EBI- Instituto Europeu de Bioinformática

G ï Guanina

ICTV ï International Committee on Taxonomy of Viruses

KEEGï Kyoto Encyclopedia of Genes and Genomes

NBRF ï NationalBiomedical Research Foundation

NCBI-Centro Nacional de Informação Biotecnológica dos Estados Unidosda América

ORF ï Open Reading Frame

PIR- Protein Information Resource

ARN ï Ácido ribonucleico

mARN - ARN mensageiro

tARN - ARN de transferência

rARN - ARNribossómico

SIB ï Swiss Institute of Bioinformatics

TrEMBL- Translated EMBL

T - Timina

UniProt- United Protein Database

VIH ï Vírus da Imunodeficiência Humana

I. INTRODUÇÃO

1.1 Problemática em estudo

Na década de 1940 foi inventado o primeiro computador. Atribuiu-se-lhe o nome de Digital, pois os dados eram armazenados como um alfabeto binário, nomeadamente,o zero e o um.A informação construía-seligando e desligando estes números.Ainda nessa mesma década, um pouco antes de 1944, Avery e os seus colaboradores (Ghose, 2010)descobriram que o ADN era a substância que carregava a informação genética.O computador digital e a descoberta de Avery surgiram em simultâneo.

A descoberta da dupla hélice no ADN, em 1953(Núñez, 1999), veio provar que a informação genética também é escrita como um alfabeto, não binário como o computador digital, mas quaternário, pois são utilizadas 4 bases azotadas, A (adenina), C (citosina), T (Timina), G (guanina). Posteriormente, percebeu-se que ambos funcionavam de forma similar, os genes podem ser ñligadosò ou ñdesligadosò, assim como no computador digital. Esta observação era suficiente para o entendimento inerente ao enquadramento da década de 50. Curioso é constatar que um dia a informática e a biologia juntar-se-iam originando uma nova área de conhecimento: a bioinformática.

Todavia, só em meados dos anos 90é que a bioinformática surge, isto porque apesar da estrutura do ADN ter sido desvendada em 1953, nesta época a informação que contida não era lidae, à data da referida descoberta, nãoexistia a tecnologia suficiente para decifrar a estrutura doADN. Acrescente-se que foi necessário, por parte da informática uma evolução considerável em áreas como, a capacidade de armazenamento de um computador, a velocidade de processamento da informação eo custo mais acessível. Na década de 70,a capacidade de armazenamento media-se em kilobytes, o que corresponde aproximadamente a 1000[1] letras.Nesta linha de pensamento, facilmente se constata que o computador dessa época não seria capaz de processar a informação de um genoma.A título de exemplo, o processamento de um novo gene (com cerca de 12.000 pares de bases ou *letras químicas*) atualmente consegue ser decifrado em um minuto. Há três anos levava 20 minutos. Há 20 anos era um trabalho que ocupava um

[1] Esta comparação está referida ao código UTF-8.

ano. É de considerar que se fosse estudado todo o genoma humano-que com os seus 3 biliões de pares de bases pode atingir uma extensão de 800 Bíblias[2]- seria impossível comporta-lo numa estrutura desta morfologia.

Por outro lado, o estudo dos vírus tem levado a desenvolvimentos científicos muito significativos nas áreas da biologia molecular e da genética. Estes, fisicamente, podem ser classificados como material genético que consegue preservar-se, iludindo as estratégias defensivas dos hospedeiros que usam para se propagar. A importância científica do seu estudo reside em dois pontos principais: no aumento do conhecimento da forma como utilizam os recursos energéticos das células que utilizam como hospedeiros e, no esclarecimento de muitos dos mecanismos biológicos dos próprios hospedeiros. As modernas biotecnologias incluem, em grande parte, o uso de agentes virais. Pode ser usado diretamente o material genético isolado, mas também, indiretamente pode obter-se resultados a partir dos estudos desenvolvidos, em especial das suas estratégias moleculares de sobrevivência.É, neste contexto relevante o facto de este assunto despertar um interesse científico óbvio, no qual as doenças causadas por vírus contribuem de forma significativa para a morbilidade e mortalidade de muitas espécies vivas, com particular realce para o seu impacto em muitas das atividades económicas e industrializadas da sociedade atual, como ocaso da indústria farmacêutica.

1.2. Objetivo

O objetivo da realização deste trabalho foi tentar compreender o estado da arte da Virologia, facilitado pela bioinformática e as suas potencialidades, mais concretamente,levar a cabo a identificaçãodas aplicações desenvolvidas dentro da bioinformática para conhecer melhor os vírus.

1.3. Metodologia

Esta pesquisa, em consonância com o seu objetivo, baseia-se na revisão da literatura científica, nomeadamente da literatura sobre a bioinformática ea bioinformática associada a Virologia.

[2]http://www.escuelapedia.com/genoma-humano-en-que-consiste/ consultado 04/08/2014

II. ENQUADRAMENTO TEÓRICO

Quando se associa bioinformática à Virologia, a nossa atenção debruça-se inevitavelmente sobre a biologia molecular, e esta leva-nos à compreensão das estruturas doADN e ARN, do vírus, do genoma eda sua sequenciação.

Todos os organismos vivos dependem dos seguintes tipos de moléculas para todas as suas funções biológicas: ácidos nucleicos,proteínas, lípidos eglícidos.Nomeadamente, os ácidos nucleicos, ADN e ARN são macromoléculasque codificam o conjunto completo de óinstruçõesô(o genoma) que são necessárias para unir, manter e reproduzir cada organismo vivo.

2.1 ADN

Desde Avery em 1944, que se conhece a estrutura básica do ADN. OADN é uma molécula composta de fosfato, desoxirribose e pelas bases adenina, guanina, citosinaetimina.No ADN a citosina emparelha com a guanina e a adenina emparelha com a timina.A espinha dorsal da molécula de ADN é uma cadeia de repetições de unidades de desoxirribose-fosfato. Éformada por duas cadeias na forma de uma dupla hélice. A dupla hélice é um fator essencial na replicação do ADN, onde cada hélice serve de molde para a nova cadeia. Assim, cada cadeia contém a informação genética codificada das caraterísticas hereditárias(Bryce & Pacini, 1994).

Nas células dos eucarióticos, o ADN está dentro do núcleo, mas nas células dos procarióticos, como estes não têm núcleo definido, o ADN está aglomerado no nucleoide.

2.2ARN

O ARN nos procarióticos é produzido no citoplasma e nos eucarióticos é produzido dentro do núcleo.

A molécula de ARN é formada por fosfato, ribose, adenina, guanina, citosina e uracilo. Durante a replicação do ARN a citosina emparelha com guanina e a adenina emparelha com o uracilo. A espinha dorsal da molécula de ARN é uma cadeia de repetição de unidades de ribose-fosfato(Bryce & Pacini, 1994).

Os três tipos principais de ARN: mARN, tARN e rARN são todos eles de cadeia simples. O mARN tem como primeira responsabilidade ser mensageiro da informação genética presente no ADN, ou seja, carrega o protótipo (projeto de estrutura) de uma proteína, desde o ADN da célula para os ribossomas, onde é produzida a proteína. O tARN leva o aminoácido apropriado para dentro do ribossoma para ser produzida uma nova proteína. Enquanto os próprios ribossomas são constituídos sobretudo de rARN.

Outros tipos de ARN fazem muito mais do que participar na síntese de proteínas:

- ARNs envolvidos na modificação da transcrição tardia ou na replicação do ADN.
  Exemplos:

  Small nuclear RNA para emendar e outras funções (Thore, Mayer, Sauter, Weeks, & Suck, 2003).

  Ribonuclease P para o desenvolvimento do tARN(Pannucci, Haas, Hall, Harris, & Brown, 1999)

- ARNs regulatórios

  Exemplos:

  MicroARN para a regulação do gene da maioria dos eucarióticos(Lin, Miller, & Ying, 2006)

  Crispr ARN para resistir aos parasitas, provavelmente atingindo o seu ADN (Brouns et al., 2008)

- ARNs parasitário

  Exemplos:

  Retrotransposon para a autopropagação nos eucarióticos e algumas baterias (Boeke, 2003)

- Outros ARNs

  Exemplos

  Vault ARN provavelmente para a expulsão dos xenobióticos(Gopinath, Matsugami, Katahira, & Kumar, 2005).

As moléculas de ARN que não assumem a forma de mARN são conhecidos como não-codificantes (ncARN).Por exemplo, muitos tipos de ARN são catalisadores, mais concretamente, estes executam reações bioquímicas como as enzimas. Outros tipos de ARN desempenham funções complexas de regulação nas células, nas quais desempenham numerosas funções, tanto nos processos normais das células como nas células doentes(Mattick, 2001).

2.3 O Genoma

O termo Genoma foi criado em 1920, por Hans Winkler, um professor da Universidade de Hamburgo. Este permitiu a definiçãodomaterial genético de um organismo.O genoma é a soma de genes que define como se vai desenvolver e funcionar um ser vivo.O genoma é transmitido de geração em geração e determina a espécie do ser vivo, porque nele encontram-se gravadas as caraterísticas hereditárias encarregues de orientar o desenvolvimento biológico de cada indivíduo(Ridley, 2000).

Pode ser aplicado especificamente para definir o que está armazenado num set completo de ADN nuclear ou ógenoma nuclearô como também é aplicado ao que está armazenado, dentro de organelos que contém o seu próprio ADN.No caso em análise, o ógenoma mitocondrialô ou o ógenoma do cloroplastoô. Destaque-se ainda que, os vírus, os plasmídeos e outros elementos transportáveiscomo o transposão contem genoma(Madigan, 2005).Mas, a maior diversidade de genomas pode ser encontrada nos vírus (University of Cape Town, 2008).

2.4 A sequenciação

A sequenciação do ADN é uma série de processos bioquímicos que tem por finalidade a determinação da ordem dos nucleótidos, cujasletras químicas ou bases são A, G, C, T. Nos anos 70 foi desenvolvida uma metodologia de sequenciação que consiste na adição de nucleótidos modificados aos quais podemos chamar dedidesoxinucleotídeos. Estes impedem o crescimento de um fragmento de ADN em replicação pelaADN polimerase, após a sua adição. Na década de 80, desenvolveu-se uma técnica rápida de sequenciação por meio da quebra de uma cadeia de ADN, sendo os fragmentos visualizados através do processo de eletroforese[3]. Poucos anos depois houve um novo avanço tecnológico por meio da introdução da técnica de interrupção da sequência, através da incorporação ao acaso dum nucleótico modificado. A referida técnica chamada de técnica de didesoxi, rapidamente tomou o lugar da anterior, possibilitando o desenvolvimento de sequenciadores automáticos de ADN- técnica conhecida por metodologia de Sanger.

A sequenciação realiza-se sob a forma de *genome projects,*pela complexidade, pelo custo e pela duração requerida.

[3] A eletroforese é uma técnica de dissociação eletrolítica na qual as partículas carregadas movem-se pela influência de forças electroestáticas para um elétrodo de carga oposta, quando é aplicada uma diferença de potencial numa solução que contem eletrólitos.

Há*genome projects*que podem ser de curto e de longo alcance. Nos *genome projects* de curto alcance, todo o ADN de uma fonte (pode ser um organismo simples como vírus, uma bactéria ou até mesmo um mamífero) é dividido em milhões de pequenos pedaços. Estes pedaços são lidos por máquinas de sequências,de forma automática, que podem ler até 1000 bases ao mesmo tempo, resultando em grandes quantidades de informação. Quando termina a leitura da sequenciação, esta informação é processada por um algoritmoque junta todos os pedaços, que deteta onde as duas das sequências curtas se sobrepõem. Refira-seque este mecanismoé de acerto e correção.(Pevsner, 2009).

Existe uma grande diferençaentre o genoma dos procariontes e dos eucariontes. Nos procariontes, a maior parte do genoma (85-90%) é não-repetitivo, quer esta definição dizer que está formado por código ADN, enquanto as regiões sem código são muito pequenas (Koonin & Wolf, 2010). No caso dos eucariontes, a variação do conteúdo repetitivo de ADN é extremamente alto, daí que o genoma das plantas e dos mamíferos seja composto de ADN repetitivo (Lewin, 2004).

Em 1976, quando Walter Fieres estabeleceu a sequência completa de nucleoide, de um genoma-ARN viral (Bacteriófago MS2) (Fiers et al., 1976), a tecnologia era muito limitada. Na atualidade, o desenvolvimento tecnológico tem facilitado exponencialmentea sequenciação que pode ser realizada a custos acessíveis.

Consequentemente, o número de genomas sequenciaisestá a aumentar rapidamente e também as bases de dados necessárias para o seu estudo. Esta multiplicação de informação tem vindo a promover a criação de centros de armazenamento e processamento de dados, patrocinados pelos governos, como por exemplo, o NCBI criado em 1988. Estepertence ao governo dos Estados Unidos de América(NCBI, n.d.).Estas imensas quantidades de informação têm vindo a promover o desenvolvimento de uma nova disciplina, a bioinformática, que pretende dar resposta à necessidade de armazenamento e tratamento sistemático, confiável dos dados.

2.5 O tamanho do genoma

Quando falamos da bioinformática, referimo-nosà gestão de grandes quantidades de dados. Isto é deve-se sobretudo aotamanho do genoma de cada ser vivo. Um genoma corresponde ao número total de pares de bases numa cópia de um genoma haploide[4]. O genoma está constituído por segmentos que podem ou não ser repetitivos, como foi explicado anteriormente. Os tamanhos de genoma são muito variados. Os procariontes e os eucariontes mais básicos possuem muito ADN não-repetitivo, enquanto os eucariontes mais evoluídos tendem a ter mais ADN repetitivo (Lewin, 2004). Na Tabela 1 é possível corroboraras grandes diferenças:

Tabela 1. Exemplos de tamanhos de genoma.

Tipo de organismo	Organismo	Tamanho (bp)	Descrição
Ameboide	Polychaosdubium ("Amoeba" dubia)	670Gb	Genoma mais comprido, mas debatido (Parfrey, Lahr, & Katz, 2008)
Planta	Paris japonica (Japanese-native)	150Gb	O mais comprido genoma de planta conhecido (Pellicer, Fay, &Leitch, 2010)
Peixe	Protopterusaethiopicus (marbledlungfish)	130Gb	O mais comprido genoma de vertebrado (Dufresne&Jeffery, 2011)
Mamífero	Homo sapiens	3.2Gb	Sequenciação do genoma humano (U.S. D.O.E &NationalInstitutesofHealth, n.d.)
Planta	Populustrichocarpa	480Mb	Primeira sequenciação de uma árvore (Set/2006) (Wilkins, Nahal, Foong, Provart, &Campbell, 2009)
Peixe	Tetraodon nigroviridis (type of puffer fish)	390Mb	O mais pequeno genoma de vertebrado estimado (BroadInstitute, n.d.)
Inseto	Bombyxmori (bicho da seda)	432Mb	14,623 genes estimados (Consortium, 2008)
Planta	Arabidopsisthaliana	157Mb	Primeiro genoma sequenciado de uma planta (2000) (Greilhuberet al., 2006)
Nematoide	Caenorhabditiselegans	100Mb	Primeiro genoma sequenciado de um animal multicelular (1998) (Initiative, 2000)
Planta	Genliseamargaretae	63Mb	O mais pequeno genoma de uma planta florida (2006) (Greilhuberet al., 2006)

[4] Haploide: refere a uma cópia do genoma em uma célula determinada. Os cromossomas são geralmente diploide.

Tipo de organismo	Organismo	Tamanho (bp)	Descrição
Fungo	Aspergillusnidulans	30Mb	(Galaganet al., 2005)
Nematoide	Pratylenchuscoffeae	20Mb	O mais pequeno genoma animal conhecido (ñAnimal Genoma SizeDatabase,ò n.d.)
Levadura	Saccharomycescerevisiae	12.1Mb	O primeiro genoma de eucarionte sequenciado (1996) (ñSaccharosmyces Genome Database,ò n.d.)
Vírus	Pandoravirussalinus	2.47Mb	O mais comprido genoma (Philippe et al., 2013)
Bactéria	Haemophilusinfluenzae	1.8Mb	O primeiro genoma de um organismo vivo sequenciado (1995) (Fleischmannet al., 1995)
Vírus	Megavirus	1.3Mb	O mais comprido genoma viral conhecido até 2013 (Legendre, Arslan, Abergel, &Claverie, 2012)
Bactéria	Nasuiadeltocephalinicola (strain NAS-ALF)	112kb	O mais pequeno genoma não-viral (Greilhuberet al., 2006)
Vírus	Phage ɔ-	48kb	Utilizado como vetor para a clonagem da recombinação do AND (Thomason, Sawitzke, Li, Costantino, & Court, 2007)
Vírus	VIH	9.7kb	(Hunt, n.d.)
Vírus	Phage ū -X174	5.4kb	A primeira sequência de genoma-AND (Ziffet al., 1973)
Vírus	Bacteriophage MS2	3.5kb	A primeira sequência de genoma-ARN (Fierset al., 1976)
Vírus	Porcinecircovirustype 1	1.8kb	O mais pequeno vírus a replicar autonomamente nas células eucariontes (Mankertz, 2008)

Fonte: Adaptado da tabela http://en.wikipedia.org/wiki/Genome (consultado 14-8-2014).

2.6 Vírus

Uma das áreas de relevância na bioinformática é o estudo dos vírus, pelo facto de serem uma das estruturas mais elementares da vida. Paradoxalmente são também estruturas complexas.Esta dualidade proporciona debate como pode verificar-se na literatura (Forterre, 2010). Apontar-se-ão como causaso facto destes seremagentes infeciosos considerados ameaça para a saúde. O Vírus(do latim *vírus*, "veneno" ou "toxina") apresenta genoma constituído de uma ou várias moléculas de ácido nucleico (ADN ou

ARN), as quais possuem a forma de fita simples ou dupla. Os vírus são seres simples formados basicamente por uma estrutura proteica chamada cápside, que protege o material genético. Em geral, os vírus só apresentam um tipo de material genético:ADN e ARN. Apenas no caso do citamegalovírus[5], que tem ADN e um ARNm(Wagner & Hewlett, 2004). O *virião* ou *vírion* é a partícula viral completa rodeada por uma cápsula protetora ou capsídeo e constitui a forma infeciosa do vírus. Fora do ambiente intracelular, os vírus são inertes(Fraenkel-Conrat & Singer, 1964)(Adelberg, Jawetz, & Melnick, 1998).

As proteínas que compõem a cápside são específicas para cada tipo de vírus. A cápside juntamente como ácido nucleico é denomina-se de núcleo cápside. Alguns vírus são formados apenas pelo núcleo cápside, outros no entanto, possuem um involucro externo ao núcleo cápside. Esses vírus são denominados vírus com involucro. O involucro consiste principalmente numa bicamada fosfolipídica derivada da membrana plasmática da célula hospedeira e em moléculas de proteínas virais, específicas para cada tipo de vírus, imersas nas camadas de lipídios. Alguns vírus possuem enzimas. Por exemplo o VIH tem a enzima Transcriptase Reversa que faz com que o processo de transcrição reversa seja realizado (formação de ADN a partir do ARN viral). Esse processo de formação de ADN a partir de ARN viral é denominado retrotranscrição, o que deu o nome retrovírus aos vírus que realizam esse processo. Os outros vírus que possuem ADN fazem o processo de transcrição (passagem da linguagem de ADN para ARN) e só depois a tradução (Wagner & Hewlett, 2004).

Dentro da célula, a capacidade de replicação dos vírus é surpreendente: um único vírus é capaz de multiplicar, em poucas horas, milhares de novos vírus, e infetar células que servem de hospedeiras. Osvírus são parasitas obrigatórios do interior celular e, tal situação significa que somente se reproduzem pela invasão e possessão do controle da maquinaria de autorreprodução celular do hospedeiro. Nesse mesmo lugar efetuar-se-á a síntese das proteínas dos vírus e, simultaneamente, será permitidaa existênciada multiplicação do material genético viral. Sendo assim, os vírusinfetameucariontes (organismos cujas células têm invólucro nuclear), mas também infetam procariontes (domínios bactéria e archaea). Em muitos dos casos os vírus modificam o metabolismo

[5]Citamegalovírus são herpes-vírus com elevada especificidade com o hospedeiro e que podem causar infeções no homem. É frequente encontrar-se em indivíduos com tumores e em portadores de VIH (Candeias, Stewien, & Barbosa, 1974)

da célula que parasitam, e podem provocar a sua degeneração e morte. Para isso, é preciso que o vírus inicialmente entre na célula. Os víruscapsulados aderem à parede celular e "injetam" o seu material genético no citoplasma. No caso dos vírus não capsulados, estes penetram na célula por endocitose mediada por recetores ou por viropexia (penetração direta) (Wagner & Hewlett, 2004).

Os recetores das células são os que determinam qual tipo de vírus irá infetar a célula. Geralmente, o grupo de células que um tipo de vírus infecta, é bastante restrito. Existem vírus que infetam apenas bactérias, denominadas de bacteriófagos; os que infetam apenas fungos,denominam-se demicófagos; os que infetam as plantas e os que infetam os animais, denominados, respetivamente, vírus de plantas e vírus de animais(Flint, Racaniello, Enquist, & Skalka, 2009).

2.6.1 Classificação dos vírus

Os vírus representam a maior diversidade biológica do planeta, sendo mais diversos que as bactérias, as plantas, os fungos e os animais juntos(Cann, 2001). Tal situação traduz-se num gigantesco número de genes. Sendo assim, para compreender os vírus, tem vindo a ser necessária a organização de grandes quantidades de informação viral, resultandoesta nacriação de uma diversidade de classificações.

Inicialmente, os vírus foram denominados de acordo com o nome da doença que ocasionavam. A primeira tentativa para classificar os vírus foi proposta por Johnson em 1927, uma nomenclatura baseada em hospedeiro e prioridade de constatação, expressa em um número, conforme a ordem de sua descoberta (ex. tabaco virus1, tabaco virus2).A partir dessa altura, foram surgindo muitas outras, mas nenhuma obteve consenso.

2.6.1.1 Critérios para a elaboração das classificações

As classificações que têm vindo a ser desenvolvidas foram criadas com o intuito de facilitar o estudo e a compreensão da diversidade de vírus.Na atualidade, é possível agrupar a diversidade de classificações de acordo com aspropriedades e critérios considerados:

a) Morfologia do virão, incluindo tamanho, forma, tipo de simetria, presença ou ausência de espiculas ou peplômeros[6], ea presença ou ausência de invólucros.

b) Propriedades do genoma do vírus, incluindo tipos de ácido nucleico (ADN ou ARN), tamanho do genoma em kilo-bases (kb), cadeia (simples ou dupla), linear ou circular, direção (positivo, negativo), segmentos (número, tamanho), sequência do nucleótido e conteúdo G + C.

c) Propriedades físico-químicas do virão, incluindo massa molecular, estabilidade do pH, estabilidade termal, e suscetibilidade a agentes físicos e químicos, nomeadamente éter e detergentes.

d) Propriedades das proteínas dos vírus, incluindo número tamanho e atividade funcional das proteínas estruturais e não-estruturais, sequência do aminoácido, e modificações (glicosilação, fosforilação).

e) Organização e replicação do genoma, incluindo ordem do gene, número e posição da grelha de leitura aberta[7], estratégia de replicação (padrões de transcrição, traslação), localização na célula (acumulação de proteínas, encaixe do virião, tipo de virião).

f) Propriedades biológicas, incluindo gama de hospedeiros naturais, modo de transmissão, relações do vetor, patogenicidade, tropismo tecidual, e patologia (Flint et al., 2009; Wagner & Hewlett, 2004).

Vários sistemas de classificação foram criados, sendo que estes geramalguma dificuldade na gestão do conhecimento do vírus. Uma resposta a este problema foi a criação do *InternationalCommitteeonTaxonomyofViruses*.

2.6.1.2 Sistemas de classificação utilizados

a) Classificação ICTV

[6] São estruturas proeminentes, geralmente constituídas de lípidos e glicoproteínas, expostas na superfície do involucro viral de certos vírus (Cann, 2001).

[7] A grelha de leitura aberta, em Inglês *Open Reading Frame*(ORF), conhecida também por *proteincodingsequences*é uma porção de uma molécula de ADN que, quando transferida dentro dos aminoácidos, contem codões de iniciação. O código genético lê sequências de ADN em grupos de três pares de base, o que quer dizer que uma molécula de ADN de cadeia dupla pode ler-se em qualquer das seis possíveis grelhas de leitura (National Human Genome Research Institute, n.d.)

O ICTV, desde 1971, tem vindo a desenvolver o sistema atual de classificação taxonómica utilizando as propriedades mais certas dos vírus para manter a uniformidade familiar. A estrutura taxonómica geral é a seguinte:

Ordem: virais
Família: *viridae*
Subfamília: *virinae*
Gênero: vírus
Espécie: vírus

Na atual taxonomia do ICTV (versão 2013), foram estabelecidas sete ordens: Caudovirales, Herpevirales, Ligamenvirales, Mononegavirales, Nidovirales, Picoarnvirales e Tymovirales (ICTV, 2014).

b) Classificação Baltimore

O sistema de classificação taxonómica do ICTV é utilizado juntamente com o sistema de classificação Baltimore. A classificação de Baltimore(figura 1) está baseada no método de sínteses do mARN viral(Flint et al., 2009).O biólogo, premio Nobel, David Baltimore concebeu este sistema de classificação que foi apresentado à comunidade científica em 1971.Os vírus devem gerar mARNs desde os seus genomas para produzir proteínas e replicarem-se a si próprios,utilizando diferentes mecanismos para o conseguir em cada família de vírus. Os genomas virais podem ser de cadeia simples (ss) ou dupla (ds), ARN ou ADN, e podem ou não utilizar transcriptase reversa (RT)[8]. Osvírus dessARN podem também ser de cadeia positiva ou negativa. Esta classificação coloca os vírus em sete grupos:

[8] Transcriptase reversa é uma enzima usada para gerar ADN complementar (cADN), utilizado para a replicação de retrovírus (Wagner & Hewlett, 2004).

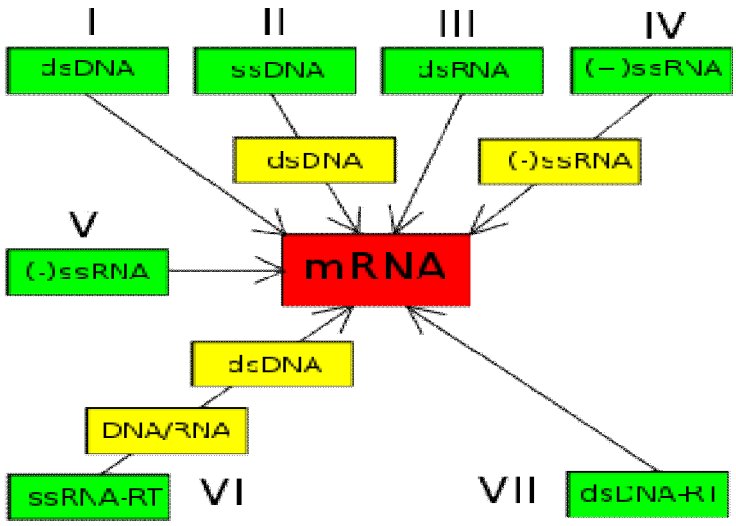

Figura 1. Classificação de Baltimore(Carter & Saunders, 2007).

c) Classificação Holmes

Holmes (1948) utilizou o sistema CarolusLinnaeus(Knapp, n.d.)para classificar os vírus em 3 grupos por uma ordem:Virales. Foram coligados da seguinte forma:

Grupo I: Phaginae (ataca bactérias)

Grupo II: Phytophaginae (ataca plantas)

Grupo III: Zoophaginae (ataca animais) (Wagner & Hewlett, 2004).

d) Sistema de Classificação de Vírus LHT

O sistema de classificação LHT está baseado em caraterísticas físicas e químicas: o ácido nucleico (ADN ou ARN), simetria (helicoidais, cúbicos ou complexos), presença de envelope, diâmetro do capsídeo, número de capsómeros. Esta classificação foi aprovada pelo ICTV em 1962(Wagner & Hewlett, 2004).

2.6.2 Os genomas dos vírus

Os vírus mostram uma grande variedade de estruturas. Dos milhões de diferentes tipos de vírus, apenas uns 5000 têm vindo a ser descritos detalhadamente(Dimmock et al., 2007).Diversos são os processos responsáveis por gerar avariabilidade genética dentro de uma população viral. Entre tais processos, estão: mutações, recombinações, rearranjos genéticos em coinfecções, entre outros. A fidelidade e a frequência dos

processos de replicação, as taxas de ocorrência de coinfecções, o modo de transmissão, o tamanho e a estrutura das populações (virais e de hospedeiros) são fatores que influenciam a geração da variabilidade genética viral. Quando os vírus se multiplicam no interior de uma célula, o material genético viral pode sofrer mutações, originando uma grande diversidade genética a partir de um único tipo de vírus. Os vírus de ARN, que dependem das enzimas ARN polimerase ou transcriptase reversa para se replicar, apresentam taxas de mutação mais elevadas, se comparados a vírus de ADN(Knipe et al., 2001), visto que as polimerases de ARN virais não têm a capacidade deprova-de-leitura[9] das polimerases do ADN.

Os genomas virais podem ser circulares, como é o caso do polyomavirus,ou lineares como acontece no adenovírus. O tipo de ácido nucleico é irrelevante na determinação da forma do genoma. Entre os vírus deARN e alguns vírus deADN, o genoma está frequentemente dividido em partes separadas, sendo este caso chamado *segmentado*. Para os vírus ARN, cada segmento frequentemente codifica só uma proteína e aparecem agrupados na cápside. No entanto, não é necessárioter todos os segmentos dentro do mesmo virião para que este seja infecioso, como no caso do *brome mosaico vírus* e alguns outros vírus de plantas (Balows & Duerden, 1998).

O Tamanho do genoma do vírus

A maioria dos vírus apresenta tamanhos muito pequenos que estão além dos limites de resolução dos microscópios óticos, sendo comum para a sua visualização o uso de microscópios eletrónicos. Nesse mundo submicroscópico, o tamanho do genoma varia grandemente entre espécies. O genoma viral mais pequeno, o ssDNAcircovirus da família Circoviridae, codifica para duas proteínas e tem um tamanho de duas kilobases(Swiss Institute of Bioinformatics, n.d.);enquanto o maior ï o pandoravirus ï tem um tamanho de aproximado de dois megabases, codificando-se para aproximadamente umas 2500 proteínasque são visíveis ao microscópio ótico (SwissInstituteofBioinformatics, n.d.-b).

Em geral, os genomas dos vírus ARN são mais pequenos que os vírus ADN porque apresentam taxas mais altas de erro na replicação e têm um limite de tamanho. Além deste limite, os erros no genoma fazem deles não viáveis ou não competitivos. Para compensar esta limitação os vírus ARN frequentemente apresentam-se segmentados

[9] Prova-de-leitura é o termo usado em genética para referir aos processos de erros de codificação (Hopfield, 1974).

para reduzir as hipóteses de que a ocorrência de um erro em um componente só, incapacite todo o genoma. Em contraste, os vírus ADN geralmente tem genomas mais compridos porque são replicados por enzimas mais fiáveis (Pressing & Reanney, 1984). O vírus ADN de uma cadeia ou ssADNpodem ser uma exceção a esta regra, conforme estudos desenvolvidos porDuffy&Holmes em geminivírus, explicada pela muito alta frequência de mutação, até ao ponto deste aproximar-se do extremo das taxas de mutação dosssARN[10](Duffy & Holmes, 2009).

III. BIOINFORMÁTICA

A bioinformática é um campo científico interdisciplinar que desenvolve métodos e ferramentas de *software* para o armazenamento, recuperação, organização e análise de dados biológicos. Sendo um campo interdisciplinar, a bioinformática combina ciência da computação, estatística, matemática e engenharia para estudar os dados e processos biológicos.

Os Sistemas de informação, tais como bases de dados e ontologias[11], são usados para armazenar e organizar os dados biológicos. Aanálise de dados biológicos para a produção de informação significativa envolve a escrita e a execução de programas de *software* que vão buscar algoritmosàteoria dos grafos[12], à inteligência artificial, àSoft Computing[13], à sondagem de dados (mining) e ao processamento de imagens e simulações computacionais. Os algoritmos, por sua vez dependem de fundamentos teóricos, como a matemática discreta, teoria decontrolo, teoria de sistemas, teoria da informação e da estatística(Hogeweg, 2011).

[10] ñA alta frequência de mutação e as altas taxas de substituição estimadas do geminivírus, as altas frequências de mutação e das taxas de substituição de nucleótidos de outras famílias de vírus ssADN, (as taxas de mutação dos fagos ØX174 e M13, aproximam-se às dos vírus do mosaico do tabaco) determinam asaltas taxas de mutaçãodos vírus dessADN(Duffy&Holmes, 2009:1545) ò.

[11] Nas ciências da informática e da informação utiliza-se o termo *ontologia* para representar o conhecimento como uma hierarquia de conceitos dentro de um domínio, utilizando um vocabulário partilhado para os tipos, propriedades e interrogações destes conceitos (Chandrasekaran, Josephson, & Benjamins, 1999)

[12] A teoria dos grafos é um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto, utilizando estruturas chamadas de grafos. (Biggs, Lloyd, & Wilson, 1986)

[13] Soft computing é um termo aplicado nas ciências da computação para se referir a problemas da informática cujas soluções são imprevisíveis, improváveis (Zadeh, 1994).

A bioinformática é similar mas distinta da biologia computacional. Esta última utiliza bioengenharia e biologia para construir bio-computadores[14], enquanto a bioinformática utiliza a computação para compreender melhor a biologia. A bioinformática e a biologia computacional perseguem os mesmos objetivos e abordagens científicas, mas diferem em escala: a bioinformática organiza e analisa os dados biológicos base enquantoa biologia computacional constrói modelos teóricos de sistemas biológicos (Waterman, 1995).

3.1 Tipos de arquivos de dados da biologia molecular

As bases de dados da biologia molecular contêm sequências de ácidos nucleicos e de proteínas, estruturas e funções de macromolécula, padrões de expressão[15], redes de vias metabólicas, e cascatas de regulação[16].

Segundo Artur Lesk (2008:29) estes arquivos de dados incluemas bases de dados primárias, as bases de dados derivadas ou secundárias, as bases de dados bibliográficos e bases de dados de sítios na web. Estas últimas bases de dados virtuais especializadas facultam funções de pesquisa conforme o objetivo, direcionadas para as necessidades de grupos de pesquisa específicos ou até de cientistas individuais.

A qualidade de uma base de dados não depende apenas da informação que contem, mas também da efetividade das suas conexões com outras fontes de informação. A importância crescente do acesso simultâneo à base de dadostem levado a pesquisa sobre a integração entre eles, sendo este o ponto em que encontrao estado da arte.

[14]Bio-computadores utilizam sistemas de moléculas biologicamente derivadas, como o ADN e as proteínas, para realizar cálculos de armazenamento, recuperação e processamento de dados. O desenvolvimento dos bio-computadores foi possível pela expansão da nanobiotecnologia. O termo nanobiotecnologia pode ser definido de múltiplas formas: como um tipo de tecnologia que utiliza materiais à escala nano, assim como materiais biológicos. Uma outra definição vê a nanobiotecnologia como o desenho e engenharia de proteínas para construir estruturas funcionais maiores. (Stix, 2001)

[15] Processo pelo qual a informação contida num gene e traduzida em estruturas presentes num determinado tipo celular (mARN ou proteínas). A acumulação exponencial de sequências genéticas tem aumentado consideravelmente à procura por metodologias, entre outras para a elucidação de padrões de expressão (National Human Genome Research Institute, n.d.)

[16] Cascata de regulação: em procariotas, é uma forma de regulação genética dos códigos de um operão para a produção de um indutor interno que gira em torno de um ou mais operões. Em eucariotas, é um modelo de múltiplas etapas de regulação genética que envolve mecanismos que fazem interface com a formação do mRNA, transporte e tradução (Parker, 1989).

3.1.1Bases de dados primários

As bases de dados primários de arquivos de informação biológica são sequências de ADN e proteínas, incluindo anotação; variações, tais como compilações de haplótipos[17], estruturas de ácidos nucleicos e proteínas, incluindo anotação; bases de dados específicas para organismos, incluindo bases de dados de genomas; bases de dados de padrões de expressão proteica; bases de dados de padrões de interação e de vias reguladoras.

O armazenamento e a organização das grandes quantidades de dados produzidos diariamente são geridos através de colaboração internacional:

O GenBank, situado no NCBI, em Bethesda, Maryland, US; o EMBLNucleotideSequenceDatabase, localizado no EBI, em Hinxton, no Reino Unido; e o Center for InformationBiology e ADNDataBank, no Instituto Nacional de Genética, em Mishima, Japão gerem em parceria o arquivo de sequências de ácidos nucleicos.

AUniProt é um consórcio entre o EBI, o SIB, e o PIR. O EBI, no Reino Unido quealoja grandes bases de dados e serviços bioinformáticos. O SIB, em Génova, mantem os servidores do ExPASy, que são os recursos centrais para as bases de dados e ferramentasda proteómica[18]. O PIR, alojado no NBRF em Washington, é o herdeiro da mais antiga base de dados de sequenciação proteica. UniProt mantem o arquivo de sequências de aminoácidos de proteínas.

Muitos projetos de sequenciação de genomas completos mantêm bases de dados focalizadas em espécies individuais como, por exemplo, o Homem, o rato, o peixe zebra, etc. Sãonotáveis o ENSEMBL do Reino Unido(Sanger Institute, n.d.) e os navegadores da Universidade da Califórnia, em Santa Cruz, Estados Unidos (CBSE, n.d.).

[17]Haplótipos são blocos de ADN transmitidos em conjunto para os descendentes. O haplótipo é um segmento do genoma que, um único cromossomo de cada par não pode ter partes da mãe e do pai juntas, apenas de um deles.São uma combinação de alelos de ligamento, quer dizer, que existem combinações de alelos em "excesso" e há uma "falta" de combinações de outros alelos. Um alelo é cada uma das várias formas alternativas do mesmo gene, ocupando um dado locus (posição) num cromossoma. Por exemplo, o gene que determina a cor da flor em várias espécies de plantas - um único gene controla a cor das pétalas, podendo haver diferentes versões desse mesmo gene. Uma dessas versões pode resultar em pétalas vermelhas, enquanto outra versão originará pétalas brancas (Watson, 1970).

[18] A proteómica é a ciência da área da biotecnologia que estuda o conjunto de proteínas e as suas isoformas, contidas numa amostra biológica, que são determinadas pelo genoma da mesma.

3.1.2 Bases de dados derivadas ou secundárias

As bases de dados derivadas ou secundárias contêm informações obtidas a partir debases primárias e das análises dos seus conteúdos. Por exemplo: motivos de sequências proteicas (ñpadrões de assinaturaò caraterísticas de famílias de proteínas); mutações e variantes nas sequências de ADN e de proteínas; classificações ou relações (conexões e caraterísticas comuns das entradas dos arquivos; por exemplo, uma base de dados de conjunto de famílias de sequências de proteínas ou uma classificação hierárquica de padrões de enrolamento de proteínas).

Em geral, as bases de dados secundárias ou derivadas estão organizadas pelo agrupamento de famílias de proteínas ou subunidades, estando baseadas no critério de similaridade entre as suas sequências.

Por exemplo, aInterpro(EMBL-EBI, n.d.) é um tipo especial de base de dados que integra os conteúdos, as caraterísticas e as anotações de diversas bases de dados individuais de famílias de proteínas, domínios e sítios funcionais.Esta basetem ainda conexões com outras bases, incluindo a classificação funcional do Gene Ontology Consortium (GO Consortium, n.d.).

IV. A VIROLOGIA NA BIOINFORMÁTICA

A virologia dentro da bioinformática tem vindo a progredir significativamente nos anos recentes. Bases de dados e ferramentas que contêm informação genómica, proteica, e funcional têm vindo a ser indispensáveis para desenvolver estudos na virologia.

A análise bioinformática nos vírus compreende tarefas relacionadas com a análise de sequências novas, incluindo identificação, anotação funcional, e análise das relações filogenéticas. Com estas ferramentas informáticas é possível responder a questões, como:

- A base de dados contém as informações do que eu preciso? Ex.: Em quais bases de dados consigo encontrar as sequências de aminoácidos de uma proteína viral específica?
- Como posso organizar as informações selecionadas de bases de dados de maneira útil? Ex.: Como posso compilar uma lista de sequências de ADN do vírus da hepatite C?

No entanto cada estudo tem desafios específicos. Por exemplo, muitos vírus têm grelhas de leitura aberta sobrepostas ou deslocações do quadro de leitura[19]. Acrescente-se que, a quantidade de recombinações frequentemente impossibilita a aplicação de uma análise filogenética clássica. Os grandes volumes e a diversidade de sequências de vírus disponíveis, requerem informação organizada e integrada em repositórios específicos de vírus. Também, as pesquisas de homogeneidade virais requerem ferramentas específicas para famílias de vírus.

Segundo Lesk, as consultas mais frequentes (Lesk, 2008:31)ocorrem porque:

1. Dada uma sequência ou fragmento de uma sequência, encontram-se sequências na base de dados que sejam similares à sequência ou fragmento?

[19] Como o ADN codificador de proteínas é dividido em codões de três bases cada, inserções e deleções podem alterar um gene a ponto de sua mensagem perder o sentido. Essas alterações são chamadas de deslocamento do quadro de leitura. Essas alterações são chamadas de deslocamento do quadro de leitura. Um codãoé uma sequência de três bases nitrogenadas de ARN mensageiro que codificam um determinado aminoácido ou que indicam o ponto de início ou fim de tradução da cadeia de ARNm. Isto significa que cada conjunto de três bases consecutivas é responsável pela codificação de um aminoácido (Calderaro et al., 2004)

2. Dada a estrutura de uma proteína, ou parte de uma estrutura proteica, encontram-se estruturas de proteínas na base de dados que sejam similares à estrutura ou parte dela?
3. Dada a sequência de uma proteína de estrutura desconhecida, encontram-se estruturas na base de dados que adotem estruturas tridimensionais similares?
4. Dada a estrutura de uma proteína, encontram-se sequências na base de dados que correspondem a estruturas similares?

Para responder a estas e outras perguntas em virologia, Yan (2008) elaborou um diagrama de fluxo (figura 2) que apresentaos passos frequentemente seguidos na análise da estrutura ï funçãoou análises dos sistemas, sendo estes dois os estudos chave para compreender as doenças virais e encontrar os agentes antivirais potenciais. Cada passo corresponde a um tipo de análise, sendo que para cada análise pode existir mais do que uma ferramenta. Um resumo das ferramentas mais frequentemente utilizadas nas análises de virologia apresenta-se na Figura 2.

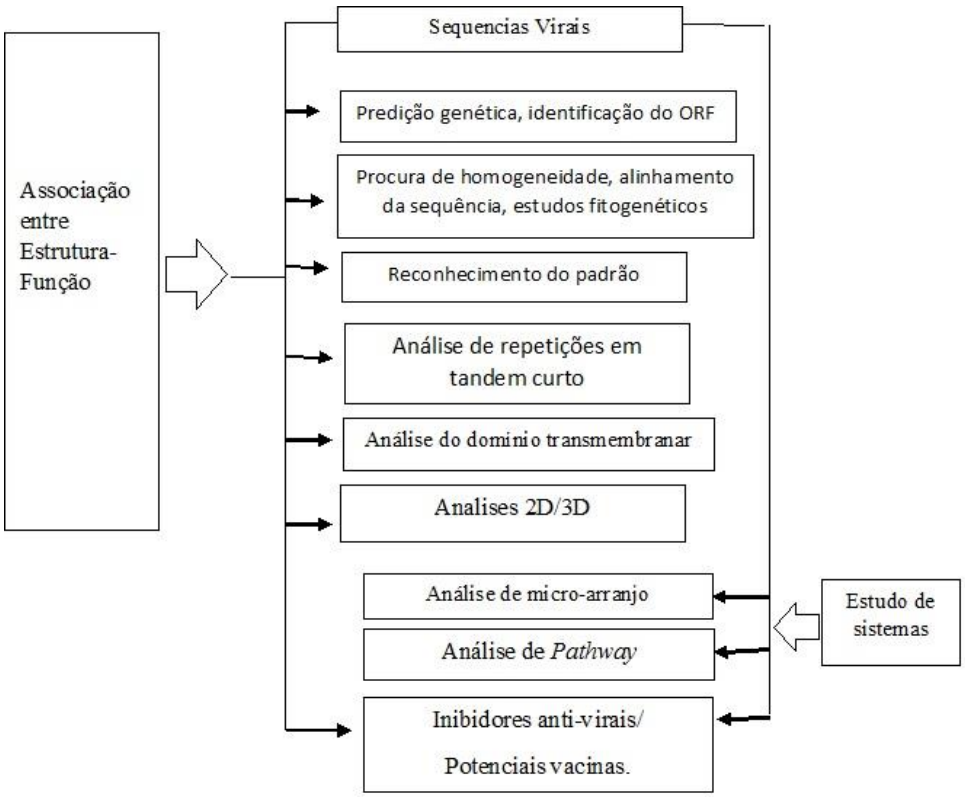

Figura 2. Diagrama de fluxo das principais análises nos estudos de virologia. Adaptadode ñThe workflow diagrama of major bioinformatic analyses in virology studiesò de Yan (2008:72).

Tabela 2. Ferramentas de análise bioinformático em estudos de virologia.

Análise	**Ferramenta**	**Link**	**Conteúdo**
Identificação do ORF	ORF Finder	http://www.ncbi.nlm.nih.gov/gorf/gorf.html	Predição de ORF para todo tipo de genoma.
	GeneMark	http://opal.biology.gatech.edu/GeneMark/genemarks.cgi	Predição para viruses
Anotação	Gene Ontology	http://www.geneontology.org	Vocabulário para anotação de genoma
Busca de homólogos	BLAST	http://www.ncbi.nlm.nih.gov/BLAST/	Busca de sequências similares
Alinhamento sequencial	ClustalW	http://www.ebi.ac.uk/clustalw/	Alinhamentos, árvores evolucionários
Programas Filogenéticos	Phylogeny Programs	http://evolution.genetics.washington.edu/phylip/software.html	Lista de programas filogenéticos
	SimPlot	http://sray.med.som.jhmi.edu/SCRoftware/simplot/	Inferir recombinação entre ancestrais de um vírus
	SplitsTree	http://www.splitstree.org	Para inferir recombinação entre ancestrais de um vírus
Reconhecimento de padrões	ScanProsite	http://www.expasy.ch/tools/scanprosite/	Predição de regiões funcionais
	Pfam	http://www.sanger.ac.uk/Software/Pfam/	Famílias de proteínas e domínios
Arquitetura modular simples	SMART	http://smart.embl-heidelberg.de/	Identificação de domínios de sinalização
Analise Epitope	IEDB	http://www.immuneepitope.org	Epítopes de anticorpos e cél. T
Repeticiones en tándem curtas	EQUICKTANDEM	http://bioweb.pasteur.fr/seqanal/interfaces/equicktandem.html	Explora uma sequência para potenciais repetições tamdem
Identificação do domínio transmembranar	Tmpred	http://www.ch.embnet.org/software/TMPRED_form.html	Predição das regiões de extensão membranar e a sua orientação
	DAS	http://www.sbc.su.se/ĕmiklos/DAS/	
	SOUSI	http://bp.nuap.nagoya-u.ac.jp/sosui/	
Predição da estructura	PredictProtein	http://www.predictprotein.org/	Predição de estrutura secundária
Modelamento da estrutura em 3D.	SWISS-MODEL	http://swissmodel.expasy.org/	Modelação de homologia 3D
Análisis da rota.	KEGG	http://www.genome.ad.jp/kegg/kegg2.html	Vias metabólicas e de regulamentação
Microarrays	EBI Microarray Informatics	http://www.ebi.ac.uk/microarray	Gestão, armazenamento e analise de microarrays
	Gene Expression Omnibus	http://www.ncbi.nih.gov/geo/	Expressão de gene y base de dados de microarrays

Fonte: Adaptado da tabela ñBioinformaticanalysistools for virologystudiesò de Yan (2008:73).

Pela diversidade e complexidade das ferramentas da virologia-bioinformática para a selecção e recuperação de sequências, considera-se pertinente seguir os critérios de Yan (2008) na sua descrição e apresentam-se aquelas que este considera como as mais comuns.

4.1 Identificação ORF e previsão de genes

O ORF é a base para procurade homólogos, análise funcional e reconhecimento das proteínas virais na identificação de potenciais agentes antivirais ou vacinas. Os ORFs podem ser identificados nas sequências de ADN ou ARN. Se um ORF é uma proteína de superfície, localizada na superfície do vírus (que pode ser verificada através dos programas de domínio transmembranar referidos mais à frente) e é particular para esse organismo, esta pode causar respostas imunes e vir a ser um candidato para uma vacina. A deteção dos ORF está no topo do diagrama de fluxo apresentado na figura 2. A sua importância está associada à criação de vacinas bacterianas, vacinas virais eaos estudos de fármacos. A identificação do ORF também é um dos primeiros passos na análise de um genoma viral desconhecido. O programa *ORF Finder*é utilizado para identificação de ORFs, sendo uma ferramenta de previsão.

4.2 Procura de homogeneidade e alinhamento desequências

A procura de homogeneidade normalmente é o passo a seguir para a anotação do genoma e para a análise funcional depois de ter realizado a pesquisa do ORF. Um alto grau de homogeneidade entre um ORF de um genoma desconhecido e uma proteína conhecida pode significar que a nova proteína tem uma função similar á já conhecida. O programa *Blast*pode ser usado na pesquisa da sequência tanto a nível dos nucleótidos, como ao nível dos aminoácidos.

Também o programa ClustalW tem sido utilizado amplamente para o estudo do genoma viral, onde os alinhamentos das sequências dos nucleótidos e aminoácido são importantes na comparação de sequências virais em diferentes espécies e cadeias. Este tipo de análise é útil para identificar semelhanças, comparar regiões conservadas e não conservadas, estabelecer relações evolutivas e construir árvores filogenéticas.

A *datamining*dos alinhamentos ClustalW permite completar uma série de comparações entre diferentes genomas virais. Por exemplo: para genomas de cadeias variadas do mesmo vírus é possível comparar regiões de sequência com menos de 90% de identidade e analisar as diferenças funcionais dentro destas regiões através da análise dos diversos padrões. Esta análise poderia fornecer pistas de como as diferenças estruturais afetam as funções, as quais podem ser um ponto indicativo para agentes antivirais potenciais ou candidatos a vacinas. Por exemplo, com o ClustalW os investigadores analisaram a data das sequenciações das variantes do *coxsackievirus A24I*, e encontraram homologias de 97% - 100% (Park et al., 2006). As árvores filogenéticas também foram construídas na base desta análise.

4.3 Reconhecimento de padrões de epítopos[20]

O padrão representa as caraterísticas comuns auma família de proteínas, constitui a região curta, mais representativa dentro de uma sequência de proteína. Os domínios de proteína de uma família em particular que vem de um antepassado comum, usualmente partilham caraterísticas funcionais. Sendo assim, a procura do padrão é um método importante para corelacionar a estrutura da sequência do genoma com a função proteômica. Os padrões funcionais podem também ser alvos potenciais para identificar inibidores antivirais.

A procura de padrões tem vindo a ser utilizada amplamente na análise de sequência de genomas. Por exemplo, a base de dados PROSITE foi utilizada na análiseda síndrome da mancha branca[21], para o estudo das funções dos produtos do gene viral. (Huang et al., 2002). No PROSITE também é possível encontrar a ferramenta PRATT, para descrever padrões num conjunto de sequênciasnão-alinhadas, o que permite encontrar padrões novos não identificados na base de dados.

[20]Epítopo ou determinante antigénico é a menor poção de antigénio com potencial de gerar a resposta imune. É a área da molécula do antigénio que se liga aos recetores celulares e aos anticorpos. São importantes para entender e compreender as doenças virais e encontrar alvos antivirais (Amabis & Martho, 2004).

[21] O vírus da Síndrome da mancha branca é o agente patogénico causador das maiores perdas económicas registadas na carcinicultura mundial. Dissemina-se rapidamente, infetando diferentes espécies de organismos aquáticos, os quais podem tornar-se reservatórios alternativos para o patogénico (Marques, 2007).

Sendo que os genomas dos vírus têm caraterísticas diferentes dos restantes organismos, e podem ter padrões especiais, em muitos casos é preciso utilizar padrões específicos para vírus. Em programas como o ScanProsite podem ser utilizados padrões definidos pelo usuário no formato PROSITE.

Uma ferramenta útil para analisar os epítopos é a *ImmuneEpitopeDatabase* que contem informação sobre anticorpos e epítopos de células T humanas, primatas não humanos, roedores e outras espécies animais. Atualmente a base de dados contem informação depurada sobre os vírus influenza, hepatites B, herpes.

4.4 Análise de repetições em tandem curto[22]

Regiões de repetição de ADN curto são sítios potenciais de mutação e variabilidade imunogénica. A deteção e comparação de mudanças em regiões de repetição de tandem curto nas sequências do genoma viral são úteis na identificação das linhagens dos vírus de distintas regiões geográficas. São vantajosas em estudos epidemiológicos de vírus patogénicos. Um motor de busca de tandem repetido muito utilizado é o Tandem RepeatsFinder. O programa procura repetições de tandem em sequências de ADN enviadas pelo usuário.

4.5 Domínio transmembranar

Os domínios transmembranar são frequentemente encontrado nas proteínas de superfície e podem ser reconhecidos pelo sistema imune. São bons candidatos a incluir nas vacinas virais. Os programas de identificação transmembranar podem predizer as regiões de extensão membranar e asua orientação para algumas sequências. Um programa utilizado é o TMpred (PredictionofTransmembraneRegionsandOrientation). Por exemplo, o TMpredjuntamente com dois outros programas de predição de domínio(SOSUI e DAS)foram utilizados na análise do vírus do mosaico do tabaco. Os

[22]Repetição em tandem curto são sequências curtas de ADN, normalmente com uma longitude entre 2 a 5 pares de base, que estão repetidas numerosas vezes com a forma de cabeça-cauda. Por exemplo: a sequência de 16 pares de bases ñgatagatagatagataò representariam 4 cópias cabeça-cauda da sequência anterior ñgataò. Os polimorfismos nas repetições de tandem curto devem-se ao número diferente de cópias do elemento de repetição que ocorrem num grupo de indivíduos (Arizona University, n.d.)

investigadores encontraram e asrecombinações do vírus do mosaico do tabaco que afetavam o tabaco mais suscetível que continham um domínio transmembranar nas subunidades de proteína de cobertura e causavam respostas necróticas (Li et al., 2006).

4.6 Estudos estruturais secundários e terciários

O desenvolvimento de modelos de uma estrutural viral pode ajudar no esclarecimento da relação estrutura-antigenicidade. Também, modelos de previsão em 3D podem, em diferentes cadeias e segmentos, ser comparados para estudar como as diferenças de alinhamento afetam a estrutura real da proteína e os sítios antigénicos, superfícies de dobramento e padrões funcionais. Estas comparações podem ser utilizadas para identificar inibidores e vacinas. Por exemplo, a previsão 3-D da estrutura da proteína foi analisada no vírus humano*syncytial respiratório* (HRSV)[23](Sugawara et al., 2002).Neste estudo observou-se diferenças estruturais significativas relacionadas com a longitude dos peptídeos que contêm a ligação cisteína (*cysteinenoose)*, encontrando-se boa correlação com a atividade imunogénica dos peptídeos.

Para desenvolver o modelo duma estrutura secundária da proteína é utilizada a ferramenta PredictProtein. Este programa permite prever a estrutura secundária e resolve a acessibilidade do solvente, e possíveis hélices transmembranar. O programa também fornece informação sobre a precisão esperada dos métodos de predição.

Para predições de estrutura 3-D, utiliza-se ferramentas como a SWISS-MODEL. Estes programas baseiam-se na busca homóloga utilizando como padrões de busca estruturas similares e conhecidas de proteínas. Estos padrões vêm sobretudo de domínios de proteínas disponíveis no Protein Data Bank. O resultado obtido inclui o padrão selecionado, a sequência de alinhamento entre a sequência de busca e o padrão, e o modelo predito.

[23] O HRSV é o vírus que causa infeções do trato respiratório. É a causa principal das infeções respiratórias inferiores. (Glezen, Taber, Frank, & Kasel, 1986)

4.7 Análise dasvias metabólicas

Através de análises bioquímicas de vias e interações proteina-proteina é possível compreender como estas moléculas interatuam entre elas, e as suas tarefas funcionais ao nível dos sistemas. Variações anormais e interações nestas vias podem contribuir para estados de doença. A informação das vias pode ser utilizada para encontrar potenciais intervenções antivirais e vacinas. Uma base de dados usualmenteutilizada para este efeito é a KEEG que fornece informação das vias metabólicas e regulatórias. Adicionalmente, contêm mapas gráficos das vias e catálogos moleculares que podem ser úteis para os virologistas.

A análise da via é determinante na compreensão dos processos de infeção viral, e a sinalização da via que permite o crescimento do vírus fornece excelentes elementos para terapias antivirais. A infeção viral pode disparar uma variedade de vias de sinalização evocando respostas antivirais da célula hospedeira. Realce-se que os vírus podem ativar-se e aproveitar as vias de sinalização celular para a sua proliferação. Portanto, as bases de dados de vias podem esclarecer os mecanismos e osprocessos da infeção viral. Por exemplo, a *apoptosissignal-regulatingkinase 1 (ASK1)*[24] e a sua via são importantes na regulação da apoptosena infeção de vírus incluindo influenza, hepatite C e outros(Sumbayev & Yasinska, 2006).ASKI é um componente da quinasses de proteína mitogénica ativada (MAP)[25], como mostra a impressão do ecrã do diagrama da via da MAP quinase(figura 3).

[24] É uma proteína-enzima conhecida também por MAP3K5. ASK1 foi encontrada envolvida no cancro, diabetes, e doenças cardiovasculares e neuro degenerativo (Hattori, Naguro, Runchel, & Ichijo, 2009).
[25] MAP são proteínas quinase específicas dos aminoácidos serina, teonina, tirosina (Hattori et al., 2009).

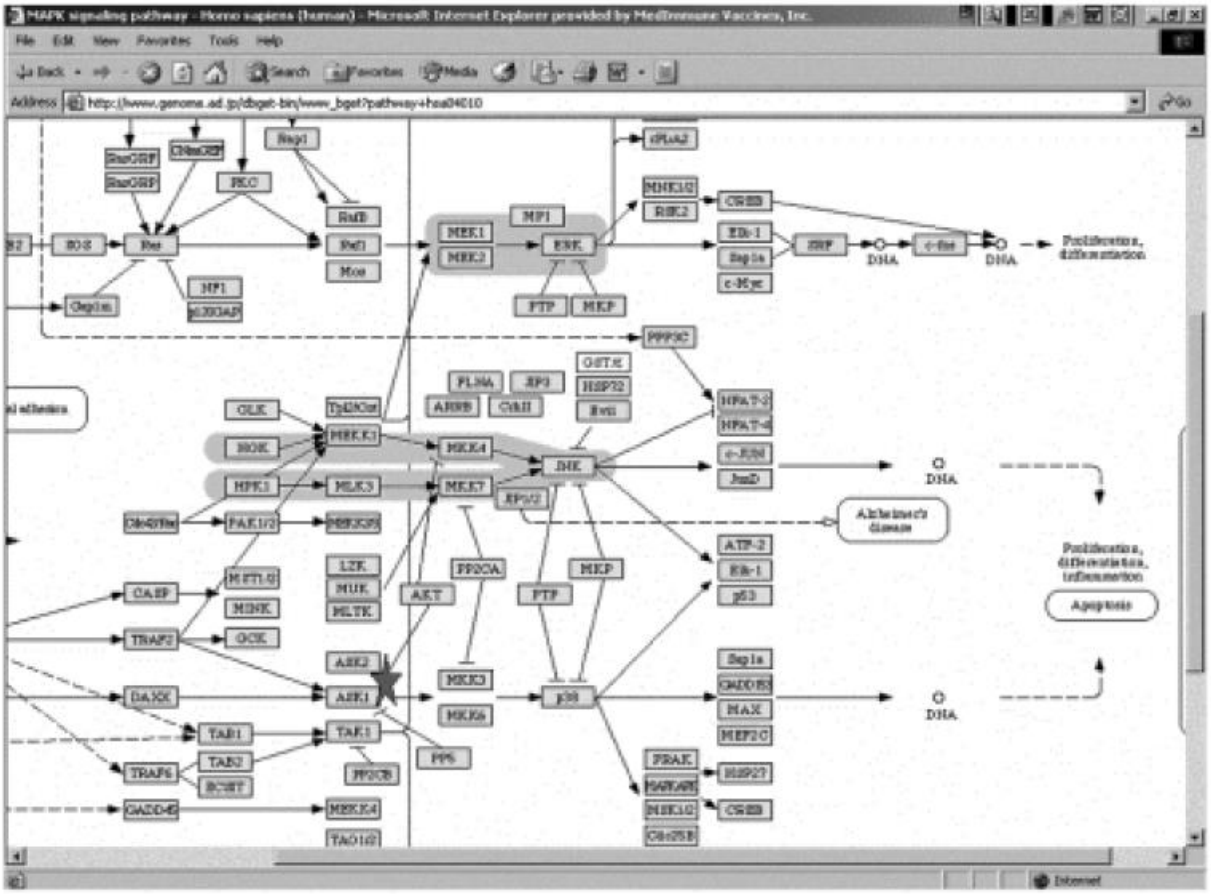

Figura 3. Impressão de ecrã da ASK1 na via da proteína quinasemitogénica ativada (MAP).Impressão tirada de um processamento de bioinformação na base de dados KEGG (Yan, 2008:80). A estrela indica a localização da ASK1.

4.8 Análise de dados de microarrays

A tecnologia para os *microarrays* permite simultaneamente examinar a expressão de milhares de genes. Esta tecnologia é útil para os estudos da genética funcional e para compreender as interações hospedeiro-vírus. A deteção de genes que se expressam de forma diferenciada conforme condições específicas pode ajudar na descoberta de mecanismos de fenómenos biológicos específicos. Os genes com padrões de expressão similar podem ser agrupados, de forma a obter pistas das vias de transdução de sinal[26]. Por exemplo, com a tecnologia *microarray*,os investigadores têm vindo a descobrir que a patogenicidade e a letalidade do vírus influenza de 1918 estãoassociadasauma resposta imune aberrante e não verificada (Park et al., 2006)

Doutra parte, existe a possibilidade dos investigadores estarem perante demasiada informação. Vários sites bioinformáticos ajudam nesta grande complexidade de manipulação de informação. O EBI, através do ArrayExpress, é um portal que fornece informação sobre manipulação, armazenamento e análise de dados*microarray*.

O Gene ExpressionOmnibus é uma base de dados de expressão genética e de *microarray*s. Este site também fornece ferramentas BLAST e outras ferramentas de apresentação.

Para terminar, existem duas fontes que convém verificar regularmente porque fornecem informação atualizada do conteúdo e das ferramentas que têm vindo a aparecer: NCBI e EBI.

[26] Transdução de sinal acontece quando uma molécula de sinalização ativa um recetor específico localizado na superfície da célula ou dentro dela. Por sua vez, este recetor dispara uma cadeia bioquímica de eventos dentro da célula, criando uma resposta. Conforme a célula, a resposta altera o metabolismo da célula, forma, expressão genética e a capacidade de divisão (Krauss, 2006)

CONCLUSÃO

Neste estudo foi possível atestar a complexidade da bioinformática, e ainda mais quando esta está associada à virologia. Mostraram-seos mecanismos e serviços científicos existentes no conhecimento da virologia, a partir da bioinformática. Esta área científica não só é complexa quanto aotema propriamente dito, mas também, este dentro da bioinformática, é o mais extenso. Fez-se uma opção pela apresentação das ferramentas, em detrimento da tentativa de descriçãodas bases de dados, sendo que a partir das primeiras é possível chegar aos dados. Apesar de muitas ferramentas terem vindo a ser desenvolvidas recentemente para estudos específicos de virologia, ainda existem limitações, conforme foi possível validar ao longo da leitura científica sobre este tema. Os novos desenvolvimentos, sejam bases de dados e/ou ferramentas, estão a crescer em função do aumento da procura.

Espera-se que esta visão resumida, mas muito complexa-desde o olhar do farmacêutico-tenha conseguido revelar a importância e utilidade dos recursos bioinformáticos disponíveis para a investigação em virologia. O foco da informação aqui analisada debruçou-se sobre as análises estrutura-função e análise de sistemas, que são essenciais para encontrar mecanismos antivirais. Outras áreas da bioinformática associadas a vírus não foram desenvolvidas porque cada uma merece um estudo por separado.

Provavelmente, a conclusão mais marcante tem sido a descoberta da profunda especialização necessária para mergulhar nesta temática, onde o grande domínio da informática e da virologia é fundamental.

Tendo sido este trabalho desenvolvido num contexto de mestrado de farmácia, fica ao alcance da profissão de farmacêutico, o conhecimento, no sentido de orientar a colocação de perguntas sobre este tema.Ocaminho a percorrer e a necessidade de ter um domínio aprofundado da virologia e da informática são ferramentas cruciais para a resposta a estas mesmas questões.

BIBLIOGRAFIA

Adelberg, E., Jawetz, E., & Melnick, J. L. (1998). *Microbiologia médica* (20th ed.). Guanabara Koogan.

Amabis, J. M., & Martho, G. R. (2004). *Biologia dos organismos*.

Balows, A., & Duerden, B. I. (1998). *Vol. 2: Systematic bacteriology*. London [etc.]: Arnold.

Biggs, N., Lloyd, E. K., & Wilson, R. J. (1986). *Graph Theory, 1736-1936*. Clarendon Press.

Boeke, J. D. (2003). The unusual phylogenetic distribution of retrotransposons: a hypothesis. *Genome Research*, *13*(9), 1975ï 1983.

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., é Van Der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, *321*(5891), 960ï 964.

Bryce, C. F., & Pacini, D. (1994). *The structure and function of nucleic acids*. Biochemical Society.

Calderaro, F. F., Doto, D. S., Baccaro, M. R., Paixão, R., Gomes, C. R., de Castro, A. F. P., & Moreno, A. M. (2004). Detecção dos genes codificadores das proteínas EF, MRP e suilisina em amostras de Streptococcus suis sorotipo 2 isoladas em suínos no Brasil. *Arquivos Do Instituto Biológico*, *71*, 15ï 19.

Candeias, J. A. N., Stewien, K. E., & Barbosa, V. (1974). Estudo sorológico de infecções ocasionadas por citomegalovírus. *Rev Saude Publica*, *8*, 257ï 263.

Cann, A. (2001). *Principles of molecular virology* (Vol. 1). Academic Press.

Carter, J., & Saunders, V. A. (2007). *Virology: principles and applications*. John Wiley & Sons.

CBSE. (n.d.). UCSC Genome Bioinformatics. Retrieved August 16, 2014, from https://genome.ucsc.edu/

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems*, *14*(1), 20ï 26.

Dimmock, N. J., Easton, A. J., Leppard, K., Dimmock, N. J., Easton, A. J., & Leppard, K. N. (2007). *Introduction to modern virology*. Blackwell Pub. Malden, MA, USA.

Duffy, S., & Holmes, E. C. (2009). Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *Journal of General Virology*, *90*(6), 1539ï 1547.

EMBL-EBI. (n.d.). InterPro: Protein sequence analasys and classification. Retrieved August 16, 2014, from http://www.ebi.ac.uk/interpro/

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., é Van den Berghe, A. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene.

Flint, S. J., Racaniello, V. R., Enquist, L. W., & Skalka, A. M. (2009). *Principles of virology, Volume 2: pathogenesis and control*. ASM press.

Forterre, P. (2010). Defining life: the virus viewpoint. *Origins of Life and Evolution of Biospheres*, *40*(2), 151ï 160.

Fraenkel-Conrat, H., & Singer, B. (1964). Reconstitution of tobacco mosaic virus IV. Inhibition by enzymes and other proteins, and use of polynucleotides. *Virology*, *23*(3), 354ï 362.

Ghose, T. (2010). Oswald Avery: the professor, DNA, and the Nobel Prize that eluded him. *Canadian Bulletin of Medical History/Bulletin Canadien D*ô*histoire de La Médecine*, *21*(1), 135ï 144.

Glezen, W. P., Taber, L. H., Frank, A. L., & Kasel, J. A. (1986). Risk of primary infection and reinfection with respiratory syncytial virus. *American Journal of Diseases of Children*, *140*(6), 543ï 546.

GO Consortium. (n.d.). Gene Ontology Consortium. Retrieved August 16, 2014, from http://geneontology.org/user-story-tags/bioinformatician

Gopinath, S. C., Matsugami, A., Katahira, M., & Kumar, P. K. (2005). Human vault-associated non-coding RNAs bind to mitoxantrone, a chemotherapeutic compound. *Nucleic Acids Research*, *33*(15), 4874ï 4881.

Hattori, K., Naguro, I., Runchel, C., & Ichijo, H. (2009). The roles of ASK family proteins in stress responses and diseases. *Cell Commun Signal*, *7*(9).

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Computational Biology*, *7*(3), e1002021.

Hopfield, J. J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proceedings of the National Academy of Sciences*, *71*(10), 4135ï 4139.

Huang, C., Zhang, X., Lin, Q., Xu, X., Hu, Z., & Hew, C.-L. (2002). Proteomic analysis of shrimp white spot syndrome viral proteins and characterization of a novel envelope protein VP466. *Molecular & Cellular Proteomics*, *1*(3), 223ï 231.

Knapp, S. (n.d.). Whatôs in a name? A history of taxonomy¡ Linnaeus and the birth of modern taxonomy,. Retrieved August 14, 2014, from http://www.nhm.ac.uk/nature-online/science-of-natural-history/taxonomy-systematics/history-taxonomy/session1/

Knipe, D. M., Howley, P. M., Griffin, D. E., Lamb, R. A., Martin, M. A., Roizman, B., & Strauss, S. E. (2001). *Fields Virology* (2nd ed., Vols. 1-2). Philadelphia: Lippincott Williams & Wilkins.

Koonin, E. V., & Wolf, Y. I. (2010). Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews Genetics*, *11*(7), 487ï 498.

Krauss, G. (2006). *Biochemistry of signal transduction and regulation*. John Wiley & Sons.

Lesk, A. M. (2008). *Introdução à bioinformática*. Artmed.

Lewin, B. (2004). *Genes VIII. 2004*. Prentice Hall.

Li, Q., Li, M., Jiang, L., Zhang, Q., Song, R., & Xu, Z. (2006). TMV recombinants encoding fused foreign transmembrane domains to the CP subunit caused local necrotic response on susceptible tobacco. *Virology*, *348*(2), 253ï 259.

Lin, S.-L., Miller, J. D., & Ying, S.-Y. (2006). Intronic microRNA (miRNA). *BioMed Research International*, *2006*.

Madigan, M. T. (2004). *Brock Biology of Microorganisms, 11th edn*. SciELO Espana.

Madigan, M. T. (2005). *Brock Biology of Microorganisms, 11th edn*. SciELO Espana.

Marques, J. da S. (2007). Contribuição para o monitoramento do vírus da Síndrome da Mancha Branca na carcinicultura de Santa Catarina.

Mattick, J. S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Reports*, *2*(11), 986ï 991.

National Human Genome Research Institute. (n.d.). Talking glossary of genetic terms. Retrieved July 31, 2014, from https://www.genome.gov/Glossary/

NCBI. (n.d.). Genome [National Center for Biotechnology Information]. Retrieved August 12, 2014, from http://www.ncbi.nlm.nih.gov/genome?itool=toolbar

Núñez, R. (1999). El siglo de la ciencia. *Muy Especial*, *42*.

Pannucci, J. A., Haas, E. S., Hall, T. A., Harris, J. K., & Brown, J. W. (1999). RNase P RNAs from some Archaea are catalytically active. *Proceedings of the National Academy of Sciences*, *96*(14), 7803ï 7808.

Park, K., Lee, K., Lee, J., Yeo, S., Lee, S., Cheon, D., é Jee, Y. (2006). Acute hemorrhagic conjunctivitis epidemic caused by coxsackievirus A24 variants in Korea during 2002ï 2003. *Journal of Medical Virology*, *78*(1), 91ï 97.

Parker, S. P. (1989). McGraw-Hill dictionary of scientific and technical terms.

Pevsner, J. (2009). *Bioinformatics and functional genomics*. John Wiley & Sons.

Pressing, J., & Reanney, D. C. (1984). Divided genomes and intrinsic noise. *Journal of Molecular Evolution*, *20*(2), 135ï 146.

Ridley, M. (2000). *Genome*. Harper and Collins.

Sanger Institute. (n.d.). Ensembl Genome Browser. Retrieved August 16, 2014, from http://www.ensembl.org/

Stix, G. (2001). Little big science. *Scientific American*, *285*(3), 26ï 31.

Sugawara, M., Czaplicki, J., Ferrage, J., Milon, A., Haeuw, J.-F., Power, U. F., é Beck, A. (2002). Structureī antigenicity relationship studies of the central conserved region of human respiratory syncytial virus protein G. *The Journal of Peptide Research*, *60*(5), 271ï 282.

Sumbayev, V. V., & Yasinska, I. M. (2006). Role of MAP Kinase-Dependent Apoptotic Pathway in Innate Immune Responses and Viral Infection. *Scandinavian Journal of Immunology*, *63*(6), 391ï 400.

Swiss Institute of Bioinformatics. (n.d.). Circovirus. Retrieved November 11, 2014, from http://viralzone.expasy.org/all_by_species/118.html

Thore, S., Mayer, C., Sauter, C., Weeks, S., & Suck, D. (2003). Crystal Structures of the Pyrococcus abyssi Sm Core and Its Complex with RNA COMMON FEATURES OF RNA BINDING IN ARCHAEA AND EUKARYA. *Journal of Biological Chemistry*, *278*(2), 1239ï 1247.

University of Cape Town. (2008). introductory microbiology (MCB2016F) and a 20-lecture course (MCB3024S, Defence and Disease). Retrieved October 8, 2014, from http://www.mcb.uct.ac.za/tutorial/virtut2.html#Note

Wagner, E. K., & Hewlett, M. J. (2004). *Basic Virology*. Blackwell Science. Retrieved from http://books.google.pt/books?id=3T6P7wUByugC

Waterman, M. S. (1995). *Introduction to computational biology: maps, sequences and genomes*. CRC Press.

Watson, J. D. (1970). Molecular biology of the gene. *Molecular Biology of the Gene.*, (2nd edn).

Yan, Q. (2008). Bioinformatics databases and tools in virology research: an overview. *In Silico Biology*, *8*(2), 71ï 85.

Zadeh, L. A. (1994). Fuzzy logic, neural networks, and soft computing. *Communications of the ACM*, *37*(3), 77ï 84.