An Analysis of Adenovirus Genomes Using Whole Genome Software Tools

Padmanabhan Mahadevan*

Department of Biology, University of Tampa, 401 W. Kennedy Blvd. Box 3F, Tampa, FL 33606; Padmanabhan Mahadevan - E-mail: pmahadevan@ut.edu; *Corresponding author

Received July 8, 2016; Revised July 28, 2016; Accepted July 29, 2016; Published August 30, 2016

Abstract
The evolution of sequencing technology has lead to an enormous increase in the number of genomes that have been sequenced. This is especially true in the field of virus genomics. In order to extract meaningful biological information from these genomes, whole genome data mining software tools must be utilized. Hundreds of tools have been developed to analyze biological sequence data. However, only some of these tools are user-friendly to biologists. Several of these tools that have been successfully used to analyze adenovirus genomes are described here. These include Artemis, EMBOSS, pDRAW, zPicture, CoreGenes, GeneOrder, and PipMaker. These tools provide functionalities such as visualization, restriction enzyme analysis, alignment, and proteome comparisons that are extremely useful in the bioinformatics analysis of adenovirus genomes.

Key Words: Adenovirus, Genome, Software, Bioinformatics, CoreGenes, GeneOrder, Percent identity, Restriction enzyme, Genomics

Background:
Human adenoviruses (HAdVs) were first discovered in human adenoid tissue in the 1950s [1]. Since then, many different HAdVs have been identified. HAdVs, like all adenoviruses, possess double stranded DNA genomes [2]. The size of the HAdV genome is approximately 35 kb. There are seven species of HAdVs (A through G) and each species consists of different HAdV types. HAdVs cause many diseases such as respiratory disease, conjunctivitis, and gastroenteritis. HAdVs are classified based on several criteria including serum neutralization assays, restriction enzyme analysis (REA), hemagglutination, phylogenetic analysis, and whole genome analysis [3].

The improvement of genome sequencing technology has revolutionized the field of genomics and this impact has certainly been felt on adenovirus genomics. The number of HAdV genomes that have been sequenced has increased at an incredible rate. Bioinformatics analysis can be applied to whole genomes in order to distinguish between HAdVs and gain insight into their evolution [4]. Indeed, whole genome sequence analysis has emerged as the gold standard for the classification of HAdVs [5].

In order to perform whole genome bioinformatics analysis on HAdVs, the appropriate whole genome software tools must be used. These myriad tools vary from standalone to web-based programs. Some of these tools include Artemis, EMBOSS, pDRAW, zPicture, CoreGenes, GeneOrder, and PipMaker (Table 1). The use of these whole genome tools in describing the relationships between HAdVs is presented here. In addition, the use of whole genome percent identities and the use of inverted terminal repeats (ITRs) as techniques to complement these tools and to describe the relatedness between HAdVs are also explored.

Methodology:
Artemis
Artemis is a genome browser that can be used to annotate genomes [6]. It can be downloaded from http://www.sanger.ac.uk/resources/software/artemis/. In addition to annotation, Artemis can be used to view and compare annotated genomes. These genomes can be downloaded from GenBank (www.ncbi.nlm.nih.gov) and examined in Artemis. Of particular value is the ability to compare two HAdV genomes by opening up two instances of Artemis and laying the windows on top of each other. This technique allows the comparison of individual proteins and their corresponding nucleotide sequences. This makes it easy to spot mutations such as mis-sense and non-sense mutations. For example, the genome of the HAdV-B3 GB prototype strain is 98% identical to the HAdV-B3 NHRC 1276 field strain. Despite this high level of sequence identity, there are differences in some proteins that can be spotted using Artemis. For example, as described previously [7], a 20.6 kDa protein in the E2B region is found in HAdV-3 GB (Figure 1A), but using
Artemis, this protein can be seen to be severely truncated by 1 stop codon in the HAdV-3 NHRC 1276 field strain (Figure 1B).

Table 1: The names and locations of the various bioinformatics tools used to analyze adenovirus genomes.

| Tool            | Location                                      |
|-----------------|-----------------------------------------------|
| Artemis         | [http://www.sanger.ac.uk/science/tools/artemis](http://www.sanger.ac.uk/science/tools/artemis) |
| % identity analysis (EMBOSS) | [http://emboss.sourceforge.net/](http://emboss.sourceforge.net/) |
| pDRAW for virtual REA | [http://www.acaclone.com](http://www.acaclone.com) |
| Clustalo for ITR analysis | [http://www.ebi.ac.uk/Tools/msa/clustalo/](http://www.ebi.ac.uk/Tools/msa/clustalo/) |
| zPicture        | [http://zpicture.dcode.org](http://zpicture.dcode.org) |
| CoreGenes       | [http://binf.gmu.edu:8080/CoreGenes3.5](http://binf.gmu.edu:8080/CoreGenes3.5) |
| GeneOrder       | [http://binf.gmu.edu:8080/GeneOrder4.0](http://binf.gmu.edu:8080/GeneOrder4.0) |
| PipMaker        | [http://pipmaker.bx.psu.edu/pipmaker/](http://pipmaker.bx.psu.edu/pipmaker/) |

Figure 1: (A) Artemis view of a 20.6 kDa protein in HAdV-B3 GB; (B) Artemis view of the same protein in HAdV-B3 NHRC1276 truncated by a stop codon which is symbolized by the “+” sign.
Whole genome percent identity analysis

It is very useful to know how closely related adenoviruses are to each other. One way of determining this is to examine the whole genome nucleotide percent identity of an adenovirus genome and compare it to the percent identity of another adenovirus genome. This can be accomplished using the EMBOSS package [8] which contains programs that perform pairwise alignment of two sequences and output the percent identity between them. Specifically, the two programs are called needle and stretcher. Needle performs a classic Needleman-Wunsch global alignment of two sequences, while stretcher uses a modified version of the same algorithm to deal with longer sequences. An example of the utility of using percent identity in comparing HAdVs can be seen in the case of HAdV-G52 which is associated with gastroenteritis [9]. There was debate as to whether HAdV-G52 was a new type of HAdV or whether it belonged to the existing HAdV-F species which are also associated with gastroenteritis [10]. One piece of evidence that argues that HAdV-G52 is indeed a new type is whole genome nucleotide percent identity of this genome compared to the genomes of SAdV-1, SAdV-7, HAdV-F40, and HAdV-F41. These percent identities are shown in Table II. The HAdV-F40 and HAdV-F41 genomes have significantly lower percent identities when compared to SAdV-1 and SAdV-7. This suggests that HAdV-G52 is more closely related to the simian adenoviruses SAdV-1 and SAdV-7 than to HAdV-F40 and HAdV-F41.

In addition to downloading the EMBOSS package (http://emboss.sourceforge.net/), the needle and stretcher programs are also available online. Needle is available at http://www.ebi.ac.uk/Tools/psa/emboss_needle/ and stretcher is available at http://www.ebi.ac.uk/Tools/psa/emboss_stretcher/.

| HAdV type (GenBank accession #) | % identity to HAdV-G52 |
|--------------------------------|------------------------|
| HAdV-G52 (DQ923122)           | 100                    |
| SAdV-1 (NC_006879)            | 95.5                   |
| SAdV-7 (DQ792370)             | 82.9                   |
| HAdV-F40 (NC_001454)          | 69.1                   |
| HAdV-F41 (DQ315364)           | 69.2                   |

Virtual restriction enzyme analysis

REA has been used for a long time as an inexpensive and quick way to distinguish between HAdV strains belonging to a certain type. For example, twelve restriction enzymes were used to distinguish between these HAdVs. However, when a virtual REA is done on these two genomes, it can be seen that they are distinct from each other. Figure 2 shows a virtual REA gel plot produced by pDRAW using the BclI enzyme for these two genomes. Lane 2 corresponds to HAdV-B3 and lane 3 corresponds to HAdV-B3 NHRC 1276. The restriction patterns are quite different between the two strains, despite their percent identity being very high.
Figure 3: Alignment of ITRs from HAdV-G52, SAdV-1, SAdV-7, HAdV-F40, and HAdV-F41. The boxed region consists of a motif that is highly conserved in mastadenoviruses. The uppercase bold sequences correspond to NFI binding sites, the underlined sequences correspond to NFIII sites, the bold italic sequences correspond to SP1 sites, and the lowercase bold sequences correspond to ATF sites.

Figure 4: The HAdV-G52, SAdV-1, and SAdV-7 ITRs have been extended to 210 nucleotides. The second ATF binding site (lowercase bold) still appears to be only present in HAdV-F40 and HAdV-F41. In contrast, the SP1 site missing in HAdV-52 in the original alignment appears to be present in this extended alignment.
Figure 5: The HAdV-G52, SAdV-1, SAdV-7, HAdV-F40, and HAdV-F41 ITRs have been extended to 420 bp. Putative ATF sites appear in HAdV-G52, SAdV-1, and SAdV-7. The conserved TATA box of the E1A gene is also shown (TATTTA).
Inverted terminal repeats

The ITRs of adenoviruses are located at both ends of the linear double-stranded DNA genome. The ITRs are essential for viral DNA replication because they contain sequence motifs that serve as binding sites for cellular and viral proteins [13]. One sequence motif is the “core” origin of replication (ATAATATACC), which is highly conserved in mastadenoviruses. This site binds the pre-terminal protein-DNA polymerase heterodimer [14]. The analysis of ITRs can be used to distinguish between HAdV types as will be seen in the case of HAdV-G52, SAdV-1, SAdV-7, HAdV-F40, and HAdV-F41.

Figure 6: zPicture plots of HAdV-G52 vs. A) SAdV-1, B) SAdV-7, C) HAdV-F40, D) HAdV-F41. The red regions are evolutionarily conserved regions (ECRs) of at least 100 bp in length and at least 70% identity.

Figure 3 shows an alignment of the ITRs from these adenoviruses using ClustalW [15] (Please note that the ClustalW server has been replaced by ClustalO at EBI). The core origin is perfectly conserved in the ITRs (nucleotides 9-18), as shown in the boxed region. The ITRs also contain transcription factor binding sites to which cellular factors bind, which may reflect cell tropism. One of these is the NFI site (TGGAAACGTGCCAA), which is highly conserved between HAdV-G52, SAdV-1, SAdV-7, HAdV-F40, and HAdV-F41. The NFI site is identical between HAdV-52 and the simian adenoviruses. Similarly, the site is exactly the same between the two HAdV-F species adenoviruses. The NFIII site...
CoreGenes is a tool that is used to determine the “core” or common set of proteins in a set of genomes. It has previously been used in the classification of bacteriophages [19, 20] CoreGenes is implemented in the Java programming language and uses a combination of servlets and HTML to provide the required functionality [21, 22]. The CoreGenes algorithm takes GenBank accession numbers as input via a web interface. These genome files are then retrieved and the protein sequences are parsed and extracted from the files. Protein similarity analysis is performed for each protein from the query genome against the reference genome protein database using BLASTP from the WU-BLAST package. If the sequence alignments are equal to or greater than a user specified threshold BLASTP score, then that pair of proteins is stored and a consensus genome of related genes is created. These scores can be “custom-specified” by the user by entry into text fields in the CoreGenes web interface. It is available at http://binf.gmu.edu:8080/ CoreGenes3.5. If more than two accession numbers are entered, the CoreGenes3.0 algorithm proceeds in an iterative manner. The consensus genome created from the analysis of the first query genome and the reference genome is analyzed against the second query genome. A second consensus genome is created and stored, which is then analyzed against the third query genome. This process is repeated and the fourth query genome is treated in the same way. The final output is a table of related genes across all five genomes. CoreGenes also outputs unique genes between two genomes.

From the whole genome percent identity analysis, ITR alignments, and zPicture plots, there is strong evidence that HAdV-G52 is closely related to SAdV-1 and SAdV-7. In order to further investigate the relationship between HAdV-G52, SAdV-1, and SAdV-7, a whole proteome approach is undertaken here. The CoreGenes whole proteome analysis reveals that HAdV-G52 and SAdV-1 share a total of 35 proteins at a BLASTP threshold score of “75.” Figure 7 shows a partial table of shared proteins between HAdV-G52 and SAdV-1 that is produced by CoreGenes. The total number of proteins in HAdV-G52 is 36, while the total number in SAdV-1 is 35. Interestingly, a protein that is not annotated in SAdV-1, but which is found in HAdV-G52 is the U protein. This is likely an essential protein. For example, this protein may be involved in adenovirus DNA replication and RNA transcription [23]. Additional analysis using the annotation and genome visualization tool Artemis reveals that the U protein is in fact present in SAdV-1. These results suggest that HAdV-G52 and SAdV-1 are very closely related since they share all the same proteins with each other. This is consistent with whole genome percent identity analysis, ITR alignments, and zPicture plots.

CoreGenes analysis reveals that HAdV-G52 shares fewer proteins with SAdV-7 than with SAdV-1 with a total of 26 shared proteins at a BLASTP threshold score of “75.” The total number of proteins in SAdV-7 is 27. There appears to be several proteins unique to HAdV-G52 that are absent in SAdV-7. These are the E3 CR1-alpha1, E3 CR1beta1, E3 RIDalpha, E3 RIDbeta, E3 14.7 kDa, and U proteins. Further analysis using TBLASTN (BLAST a protein query against a translated nucleotide database) confirms that these proteins are missing in SAdV-7. In contrast, these proteins are all present in SAdV-1 and this indicates that HAdV-G52 is more closely related to SAdV-1 than to SAdV-7. This is sup-
ported by the genome identity between HAdV-G52 and SAdV-1 which is 95.5%. The genome identity between HAdV-G52 and SAdV-7 is significantly lower at 82.9%. As mentioned earlier in the zPicture analysis, a possible explanation for these missing proteins in SAdV-7 is artificial deletion of segments of the genome for use as a viral vector.

Table 1: Comparative Proteins of HAdV-G52 and SAdV-1

| Human adenovirus 52 | Simian adenovirus 1 |
|---------------------|---------------------|
| GI:117503036        | GI:61602119         |
| PRODUCT:EIA         | PRODUCT:EIA         |
| GI:117503037        | GI:61602120         |
| PRODUCT:EIB 19K     | PRODUCT:EIB 19K     |
| GI:117503038        | GI:61602121         |
| PRODUCT:EIB 55K     | PRODUCT:EIB 55K     |
| GI:117503039        | GI:61602087         |
| PRODUCT:IX          | PRODUCT:IX          |
| GI:117503040        | GI:61602088         |
| PRODUCT:IVa2        | PRODUCT:IVa2        |
| GI:117503041        | GI:61602089         |
| PRODUCT:tpol        | PRODUCT:tpol        |
| GI:117503042        | GI:61602090         |
| PRODUCT:tpTP        | PRODUCT:tpTP        |
| GI:117503043        | GI:61602091         |
| PRODUCT:52k         | PRODUCT:52k         |
| GI:117503044        | GI:61602092         |
| PRODUCT:tpIIIa      | PRODUCT:tpIIIa      |
| GI:117503045        | GI:61602093         |
| PRODUCT:III         | PRODUCT:III         |
| GI:117503046        | GI:61602094         |
| PRODUCT:tpVII       | PRODUCT:tpVII       |
| GI:117503047        | GI:61602095         |
| PRODUCT:V           | PRODUCT:V           |
| GI:117503048        | GI:61602096         |
| PRODUCT:tpX         | PRODUCT:tpX         |
| GI:117503049        | GI:61602097         |
| PRODUCT:tpVI        | PRODUCT:tpVI        |
| GI:124375533        | GI:61602098         |
| PRODUCT:Thexon      | PRODUCT:Thexon      |
| GI:117503051        | GI:61602099         |
| PRODUCT:protease    | PRODUCT:protease    |

GeneOrder4.0 uses the BLAST-like Alignment Tool (BLAT) [25] to perform efficient and fast “all-against-all” protein comparisons. GeneOrder4.0 also provides for the analysis of custom or proprietary data, that is, data not submitted to GenBank for one reason or another. Since GeneOrder4.0 is web-based, users do not have to download or install any software packages. Web-based access is especially useful for non-computationally based scientists such as bench-based biologists. Other user-friendly features of GeneOrder4.0 include zooming, printing, and customizing the final graphical plot. In addition, clicking on the data points on the plot leads to the popping up of new browser windows, leading to the GenBank record of the gene pairs on the plot.

GeneOrder
GeneOrder4.0 is a versatile user-friendly web-based tool developed for the analysis of gene order and synteny [24]. This software has been updated to analyze larger sized bacterial genomes of around 4-5 megabases (Mb). It performs “on-the-fly” analysis of two genomes and produces a dot plot of gene pairs.

In order to visualize the relatedness of the HAdV-G52, SAdV-1, and SAdV-7 genomes, they are analyzed as pairs using GeneOrder4.0. The plot between HAdV-G52 and SAdV-1 shows several related proteins and confirms that these two genomes are related to each other (Figure 8A). The plot between HAdV-G52 and SAdV-7 shows several related proteins (Figure 8B), but the num-

Figure 7: CoreGenes analysis of HAdV-G52 vs. SAdV-1. The partial table contains links to the entire genome in GenBank, as well as to each individual gene. This allows the user to perform additional analyses on genes of interest.

Figure 8: GeneOrder plots of HAdV-G52 vs. A) SAdV-1 and B) SAdV-7. The red circles indicate BLASTP scores ≥ 200, while the blue “x” symbols indicate BLASTP scores of ≥ 100 but < than 200.
The utility of whole genome tools such as data deluge, the development of whole genome software tools even more genomes. In order to deal with this deluge, the development of new tools must be encouraged and supported. Whole genome percent identity analysis using the program in EMBoss provides a broad overview of the similarity between adenovirus genomes, while zPicture enables the visualization of regions of high percent identity in these genomes. These two tools are useful in determining that HAdV-G52 is more related to SAdV-1 and SAdV-7 than to HAdV-F40 and HAdV-F41. The ITR analysis also agrees with the whole genome percent identity and zPicture results. REA analysis using pDRAW allows the differentiation of two HAdV types that may initially look the same, but are in fact distinct. HAdV-3 GB and the field strain HAdV-3 NHRC 1276 share a very high percent identity, but pDRAW analysis shows distinct restriction patterns that distinguish these two genomes. The whole genome visualization tool Artemis allows the viewing and inspection of these two HAdV-3 genomes. Upon closer inspection with Artemis, a 20.6 kDa protein was found to be truncated in the HAdV-3 NHRC 1276 field strain. This illustrates the use of Artemis in discovering minor differences between these two HAdV-3 genomes. CoreGenes finds the common genes between a set of up to five genomes. CoreGenes analysis reveals that HAdV-G52 shares more proteins with SAdV-1 than with SAdV-7. The missing proteins in SAdV-7 likely due to an artificial deletion are also found using the CoreGenes analysis. GeneOrder analysis also visualizes these missing proteins as a gap in the synteny plot that it produces. Similarly, PipMaker also shows gaps in the dot plot between HAdV-G52 and SAdV-7, reflecting the differences between HAdV-G52 and SAdV-7. In summary, all of these whole genome tools are invaluable in analyzing adenovirus genomes. Therefore, their development and the development of new tools must be encouraged and supported.

Acknowledgements
I would like to thank Don Seto for his helpful advice in writing this paper.

References
[1] Rowe WP et al. Proc. Soc. Exp. Biol. Med. 1953 84(3): 570. [PMID: 13134217]
[2] Davison AJ et al. J Gen Virol. 2003 84(11): 2895 [PMID: 14573794]
[3] Harrach B. Encyclopedia of Virology (Third Edition). Oxford: Academic Press. 2008 p1.
[4] Torres S et al. Viruses. 2010 2(7): 1367 [PMID: 21994684]
[5] Singh G et al. J. Virol. 2012 86(8): 4693 [PMID: 22301156]
[6] Carver T et al. Bioinformatics. 2008 24(23): 2672 [PMID: 18845581]
[7] Mahadevan P et al. Virology. 2010 397(1): 113 [PMID: 19932910]
[8] Rice P et al. Trends Genet. 2000 16(6): 276 [PMID: 10827456]
[9] Jones MS 2nd et al. J. Virol. 2007 81(11): 5978 [PMID: 17360747]
[10] de Jong JC et al. J. Virol. 2008 82(7): 3809 [PMID: 18334604]
[11] Li QG and Wadell G. J. Clin. Microbiol. 1988 26(5): 1009. [PMID: 2838500]
[12] Kim Y-J et al. J. Clin. Microbiol. 2003 41(10): 4594. [PMID: 14532188]
[13] Dán A et al. Virus Genes. 2001 22(2): 175. [PMID: 11324754]
[14] Temperley SM and Hay RT. EMBO J. 1992 11(2): 761.
[15] Thompson JD et al. Curr Protoc Bioinformatics. 2002 Chapter 2: Unit 2.3. [PMID: 18792934]
[16] Mul YM et al. J. Virol. 64(11): 5510. [PMID: 2214023]
[17] Schwartz S et al. Genome Res. 2003 13(1): 103. [PMID: 12529312]
[18] Ovcharenko I et al. Genome Res. 2004 14(3): 472. [PMID: 14993211]
[19] Lavigne R et al. BMC Microbiol. 2009 9:224. [PMID: 19857251]
[20] Mahadevan P and Seto D. Adv. Exp. Med. Biol. 2010 680:379. [PMID: 20865522]
[21] Mahadevan P et al. BMC Res Notes. 2009 2:168. [PMID: 19706165]
[22] Turner D et al. BMC Res Notes. 2013 6:140. [PMID: 23566564]
[23] Tollefson AE et al. J. Virol. 2007 81(23): 12918. [PMID: 17881437]
[24] Mahadevan P and Seto D. BMC Res Notes. 2010 3:41. [PMID: 20178631]
[25] Kent WJ. Genome Res. 2002 12(4): 656. [PMID: 11932250]
[26] Elnitski L et al. Curr Protoc Bioinformatics. 2003 Chapter 10: Unit 10.2. [PMID: 18428692]
[27] Schadt EE et al. Hum. Mol. Genet. 2010 19(R2): R227. [PMID: 20858600]

License statement: This is an Open Access article which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. This is distributed under the terms of the Creative Commons Attribution License