Brief Report

Characterisation of Gene Features in Monkey Pox Virus Suggests Molecular Reasons for Its Potential High Virulence and Pathogenicity

Wenfa Ng *

Posted Date: 16 March 2026

doi: 10.20944/preprints202603.1169.v1

Keywords: monkey pox virus; gene length; therapeutic window; pH 5; theoretical 2D protein gel; histogram analysis; symptomatic relief

 doi:10.20944/preprints202603.1169.v1

Brief Report

Characterisation of Gene Features in Monkey Pox Virus Suggests Molecular Reasons for Its Potential High Virulence and Pathogenicity

Wenfa Ng

Citizen scientist, Singapore; ngwenfa771@hotmail.com

Abstract

Monkey pox virus has caused two major outbreaks in the past 4 years. Transmission of the virus is via close person-to-person contact, which suggests that the virus is of high virulence and pathogenicity. This work reports evidence gleaned from an in-house MATLAB open reading frame finder applied to the GenBank assembled genome of monkey pox virus. Using the data in the output file of the software that includes gene sequence, gene length, gene molecular weight, amino acid sequence of gene, protein molecular weight, and isoelectric point of protein, two significant results were observed. Firstly, profiling for genes of length more than 60 base pair reveal a total of 255 candidate genes, of which more than 200 are of protein length between 20 to 80 plus amino acids. Secondly, plotting the protein isoelectric point and molecular weight in a theoretical 2D protein gel plot shows that there are clusters of viral proteins of different molecular weight at different pH values from pH 4 to 12. In particular, there is a large cluster of proteins between pH 8 and 10, which suggests that alkaline blood pH may promote monkey pox virus virulence. However, absence of many proteins in the pH range of around 5, points to a potential therapeutic window where modulation and control of blood pH at 5, may provide symptomatic relief for the patient, while awaiting the patient's immune response to degrade the viral particles in the blood.

Keywords: monkey pox virus; gene length; therapeutic window; pH 5; theoretical 2D protein gel; histogram analysis; symptomatic relief

Subject areas: biochemistry, genomics, molecular biology, bioinformatics, computational biology, biomedical engineering, medicine, physiology, internal medicine

Introduction

Monkey pox virus is a zoonotic viral disease that spreads through close person-to-person contact. This work attempts to analyze the genomic structure of monkey pox virus via an in-house MATLAB open reading frame finder software. Specifically, it is of keen interest to assess if monkey pox virus possesses 60 base pair genes, as well as how the isoelectric points and molecular weight of the proteins of monkey pox virus correlate with each other. Assembled genome of monkey pox virus downloaded for this work comes from GenBank with accession number GCA_025462475.1.

Genes or open reading frames are defined by a start codon of ATG with the end of the gene defined by conventional stop codons ('TAG', 'TAA', and 'TGA'). Software developed in this work reads through each base in the genomic sequence of monkey pox virus, and profile candidate genes of length more than 60 base pair in the virus genome. Gene length, molecular weight of the gene, amino acid sequence of the gene, protein molecular weight, and isoelectric point of protein are all calculated with the help of built-in functions in the bioinformatics toolbox of MATLAB.

Reasons for profiling genes of length more than 60 base pair is that there is a recent finding that 20 amino acid motif may be of high evolutionary significance as domains or motifs of important functions across different protein families [1]. Hence, it may be useful to profile for genes with length

 doi:10.20944/preprints202603.1169.v1

of about 60 base pair to determine if short peptide could be part of the arsenal of weapons that a virus could wage against human host cells. For example, short peptides may insert into the phospholipid bilayer of human cells, and facilitate the entry of the larger viral particles.

Materials and Methods

Assembled genome for monkey pox virus was downloaded from GenBank (Accession number: GCA_025462475.1) in FASTA format. The MATLAB open reading frame finder software first uses fastaread to read this file, followed by calculation of different variables pertinent to this genome.

First, it read through each base of this assembled genome, and uses the ATG start codon to identify the starting point of a gene. A gene is defined when the corresponding stop codon ('TAG', 'TAA', and 'TGA') are found. This process is repeated until the last base pair of the genome.

Next, the length of the gene (open reading frame) was determined via the MATLAB built-in function "length", and the molecular weight of the gene was determined using the oligoprop built-in function. In addition, the protein sequence of the gene was calculated via the "nt2aa" built-in function that translates a nucleotide sequence into an amino acid sequence using the standard codon table.

This is followed by the use of the "molweight" built-in function to calculate the molecular weight of the protein. Finally, the "isoelectric" built-in function was used to calculate the isoelectric point of the protein sequence. For both the gene and protein sequences, there are built-in checks in the software to detect unnatural nucleotide or amino acids.

Results and Discussion

The first characterisation of the genetic repertoire of monkey pox virus is assessing the distribution of gene length in the genome of the virus. Figure 1 shows the histogram distribution analysis of gene length of genes in the genome of monkey pox virus. Data shows that more than 200 of the 255 genes of the virus has a length between 63 and 253 base pair, which translates to a peptide or protein of length 20 to 80 plus amino acids. These are indeed short peptides with unknown biological effects and significance, but which warrants more detailed biochemical and physiological investigations. In contrast, monkey pox virus is generally known to possess about 197 open reading frames, which are likely to be larger gene constructs [2–4].

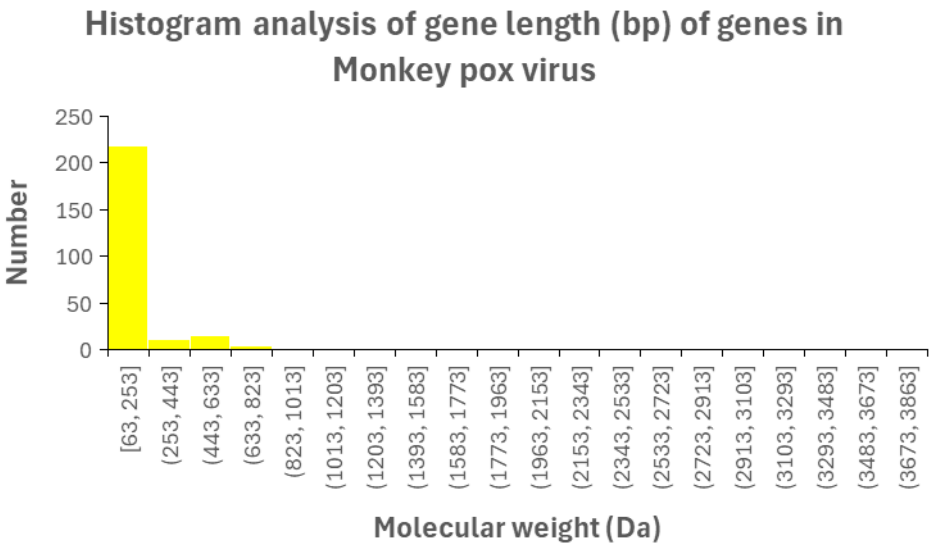

Figure 1. Histogram analysis of gene length (base pair) for genes detected in the whole genome sequence of monkey pox virus.

 doi:10.20944/preprints202603.1169.v1

Using the isoelectric point and molecular weight of protein data, it is possible to plot a theoretical 2D protein gel image for proteins in the proteome of monkey pox virus [5]. Such a theoretical plot for monkey pox virus is shown in Figure 2. One striking feature of the plot is that there is a good distribution of proteins of different molecular weight at pH 4, 6, 7, 8, 9, 10, 11, and 12. In particular, there is a dense distribution of proteins of different molecular weight between pH 8 and 10, indicating that the isoelectric point of many proteins in monkey pox virus is in the alkaline region. Such an observation suggests that monkey pox virus may exert more pathogenic effect in patients whose blood pH is in the alkaline region, above the physiological pH of 7.4.

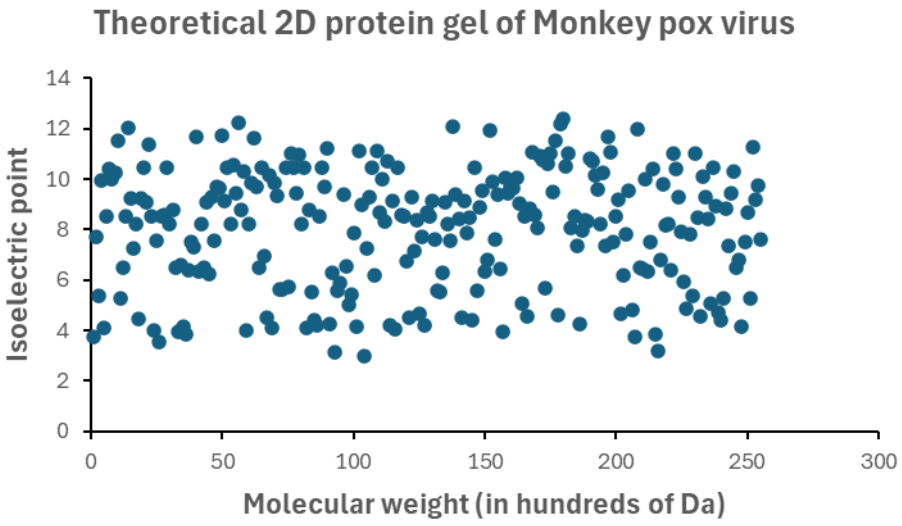

Figure 2. Theoretical 2D protein gel image for proteins in the proteome of monkey pox virus.

Medically, data in the theoretical 2D protein gel image suggests that candidate pathogenic proteins of monkey pox virus may be distributed at different pH values (isoelectric point), and the body may still suffer from severe virulence effects even if there are attempts to titrate or adjust the patients' blood pH using orally available medicines. In addition, pH 5 may be a good anchor point for adjusting patients' blood pH as there are fewer monkey pox virus proteins with isoelectric points in this pH range.

Conclusions

Clinical evidence worldwide suggests that monkey pox virus needs close person-to-person contact for transmission. This suggests that monkey pox virus may exhibit high virulence and pathogenic effect. Furthermore, high virulence means severe disease may manifest in the affected patients, which cuts the transmission chain due to severe illness that isolated the patients at home for treatment.

Such level of potential pathogenicity is evident from genomic insights gleaned by an in-house MATLAB open reading frame finder software applied to the assembled genome of monkey pox virus. Specifically, histogram analysis of the gene length of different genes in monkey pox virus shows a preponderance of short genes that encodes peptides or small proteins of 20 to 80 plus amino acids. Number of genes in this category stands at a high value of more than 200 out of 255 candidate genes.

Next, further evidence for potential high virulence and high pathogenicity of monkey pox virus comes from a theoretical 2D protein gel plot that depicts protein isoelectric point and protein molecular weight on two orthogonal axes. Specifically, in monkey pox virus, proteins of different molecular weight spans different pH values from pH 4 to pH 12, with a large cluster of proteins between pH 8 to 10. Hence, from a medical perspective, it is hard for the body to titrate the pH of

 doi:10.20944/preprints202603.1169.v1

blood in order to obliterate the pathogenic effects of a large fraction of monkey pox virus proteins. But, there appears to be few monkey pox virus proteins in the pH range of around 5, and this may offer a therapeutic window for patients with the help of orally available medicines that modulate or titrate blood pH.

MATLAB software code

```
function main



    data = fastaread('GCA_039000265.1_ASM3900026v1_genomic.fna');

    genomeseq = data.Sequence;

    genedb = genefinder(genomeseq);

    genedb_refined = gene_refine(genedb);

    genedb_table = struct2table(genedb_refined);
    writetable(genedb_table, 'Open reading frames in H1N1 virus.xlsx');


end

function genedb = genefinder(genomeseq)

    k1 = length(genomeseq);
    index = strfind(genomeseq, 'ATG');
    h = 1;
    i = 1;

    while i < k1
        if i <= k1-2
            codon_start = genomeseq(i:i+2);
        else
            codon_start = genomeseq(i:end);
        end

        flag = strcmpi(codon_start, 'ATG');

        if flag == 1
            for j = i+3:3:k1
                if j <= k1-2
                    codon_end = genomeseq(j:j+2);
                else
```

 doi:10.20944/preprints202603.1169.v1


```
                codon_end = genomeseq(j:end);
            end


            flag1 = strcmpi(codon_end, 'TAG');
            flag2 = strcmpi(codon_end, 'TAA');
            flag3 = strcmpi(codon_end, 'TGA');

            if flag1 ==0
                if flag2 ==0
                    if flag3 ==0
                        continue
                    end
                end
            end

            geneseq = genomeseq(i:j+2);
            genedb(h).geneseq = geneseq;
            h = h +1;
            i = j + 3;
            break
        end
    else
        i = i + 3;
    end



  end

end


function genedb_refined = gene_refine(genedb)

  k2 = length(genedb);
  h = 1;
  natural_aa = 'GALMFWKQESPVICYHRNDT';
  natural_nt = 'ATCGU';

  for a = 1:k2
      gene_sequence = genedb(a).geneseq;
```

 doi:10.20944/preprints202603.1169.v1


```
    k3 = length(gene_sequence);

    if k3 > 60
        genedb_refined(h).geneseq = gene_sequence;
        gene_length = length(gene_sequence);

        flag_nt = check_nt_seq(gene_sequence, natural_nt);

        if flag_nt == 1

            gene_prop = oligoprop(gene_sequence);
            gene_MW = gene_prop.MolWeight;

            protein_seq = nt2aa(gene_sequence);
        else
            gene_MW = 'Cannot be calculated';
            protein_seq = 'Contains unnatural nucleotide';
        end

        len_protein = length(protein_seq);
        protein_seq = protein_seq(1:len_protein-1);

        flag_aa = check_amino_acid_seq(protein_seq, natural_aa);

        if flag_aa == 1

            molecular_wt = molweight(protein_seq);

        else
            molecular_wt = 'Could not be calculated';
        end

        isoelectric_point = isoelectric(protein_seq);

        genedb_refined(h).gene_length = gene_length;
        genedb_refined(h).gene_Molwt = gene_MW;
        genedb_refined(h).protein_seq = protein_seq;
        genedb_refined(h).protein_molwt = molecular_wt;
        genedb_refined(h).isoelectric_pt = isoelectric_point;

        h = h + 1;
    end
end
```

 doi:10.20944/preprints202603.1169.v1


```
end


function flag = check_nt_seq(sequence,natural_nt)

    k1 = length(sequence);
    flag =1;

    for i=1:k1
        letter = sequence(i);
        if ~contains(natural_nt, letter)
            flag = 0;
            break
        end
    end

end


function flag = check_amino_acid_seq(sequence,natural_aa)

    k1 = length(sequence);
    flag =1;

    for i=1:k1
        letter = sequence(i);
        if ~contains(natural_aa, letter)
            flag = 0;
            break
        end
    end

end
```

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Dataset that supports this manuscript can be found at: Wenfa Ng (2026), "Open reading frames detected in a whole genome sequence of monkey pox virus", *figshare*, Link to dataset

Funding: No funding was used in this work.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Ng, W. Twenty Amino Acids Random Peptide Yields BLASTp Hits with Good Percent Identity. *SSRN* https://doi.org/https://dx.doi.org/10.2139/ssrn.5379485 doi:https://dx.doi.org/10.2139/ssrn.5379485.

 doi:10.20944/preprints202603.1169.v1

2. Yu, Z. et al. Genome analysis of the mpox (formerly monkeypox) virus and characterization of core/variable regions. *Genomics* **116**, 110763 (2024).
3. De Pascali, A. M. et al. Understanding the evolutionary dynamics of Monkeypox virus through less explored pathways. *Sci. Rep.* **15**, 25849 (2025).
4. Chen, S. et al. Mpox virus: virology, molecular epidemiology, and global public health challenges. *Front. Microbiol.* **Volume 16-2025**, (2025).
5. Ng, W. Theoretical 2D protein gel of Helicobacter pylori. *J. Bioinforma. Syst. Biol.* **8**, 36–37 (2025).