genes

MDPI

*Article*

# Post-Alignment Adjustment and Its Automation

**Xuhua Xia** [1,2]

[1] Department of Biology, University of Ottawa, Marie-Curie Private, Ottawa, ON K1N 9A7, Canada; xxia@uottawa.ca; Tel.: +1-613-562-5718
[2] Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, ON K1H 8M5, Canada

**Abstract:** Multiple sequence alignment (MSA) is the basis for almost all sequence comparison and molecular phylogenetic inferences. Large-scale genomic analyses are typically associated with automated progressive MSA without subsequent manual adjustment, which itself is often error-prone because of the lack of a consistent and explicit criterion. Here, I outlined several commonly encountered alignment errors that cannot be avoided by progressive MSA for nucleotide, amino acid, and codon sequences. Methods that could be automated to fix such alignment errors were then presented. I emphasized the utility of position weight matrix as a new tool for MSA refinement and illustrated its usage by refining the MSA of nucleotide and amino acid sequences. The main advantages of the position weight matrix approach include (1) its use of information from all sequences, in contrast to other commonly used methods based on pairwise alignment scores and inconsistency measures, and (2) its speedy computation, making it suitable for a large number of long viral genomic sequences.

**Keywords:** sequence alignment; automation; sum-of-pairs score; inconsistency; position weight matrix; PWM; codon-based alignment; phylogenetics



**Citation:** Xia, X. Post-Alignment Adjustment and Its Automation. *Genes* **2021**, *12*, 1809. https://doi.org/10.3390/genes12111809

Academic Editor: Sujoy Ghosh

Received: 28 October 2021
Accepted: 16 November 2021
Published: 18 November 2021



## 1. Introduction

High-quality multiple sequence alignment (MSA) is crucially important in sequence comparison and molecular phylogenetics because a poor alignment typically leads to bias and inaccuracy in phylogenetic estimation [1–3]. This is especially true in the present day, where the availability of an increasing number of sequences of increasing sequence lengths is often associated with the application of quick-and-dirty options in sequence alignment programs. This has resulted in poor sequence alignments in publications, even in prominent journals ([4,5], pp. 16–21 in reference [5]), highlighting the extent of the issue.

MSA was traditionally followed by post-alignment visual inspection and manual adjustment. However, such post-alignment improvements gradually faded away because of three contributing factors. Firstly, it becomes less important with the emergence of more accurate MSA programs such as MUSCLE [6] and MAFFT [7] with multiple iterations of MSA refinement [7–9]. Secondly, MSA in the genomic era often involve thousands of long sequences, as was frequently performed in sequence comparison of SARS-CoV-2 genomes [10,11], rendering it impractical to perform manual adjustment. Thirdly, post-alignment adjustment can be error prone and inconsistent because there is no explicit and consistent criterion that is universally used by researchers.

While sequence alignment with dynamic programming is guaranteed to generate the optimal sequence alignment given a scoring scheme [5,12], or at least one of the equally optimal alignments, progressive MSA has always been used in practice, generating sequence alignment that may well be suboptimal. This is because what is suboptimal is often not obvious when aligning closely related sequences. Only when more sequences are added to the alignment can one observe the suboptimality in previous alignment [13]. Multiple iterations of new guide trees and new MSA cannot refine such suboptimal alignment.

I illustrate two such cases. The first involves aligned sequences (Figure 1) taken from a study of mammalian ACE2 sequences in an effort to predict which mammalian species might be susceptible to SARS-CoV-2 infection [14]. The sequences were aligned with MAFFT with all optimization options selected. Only the first 25 amino acid sites are shown. We note that "-T" at sites 20 and 21 in *Nyctereutes procyonoides* should be "T-". However, *N. procyonoides* is more closely related to *Procyon lotor* and *Mustela putorius furo*, so the three sequences will be aligned first in progressive MSA. "-T" and "T-" are equally good when aligning these three sequences, so one of the two equally good alignments needs to be chosen. MAFFT happens to choose "-T", which turns out to be suboptimal when the first four sequences are added to the MSA.

```
                                1         2
                       12345678901234567890123 45
Callithrix jacchus     MSGSFWLLLSLVAVTAAQSTIEEQA
Homo sapiens           MSSSSWLLLSLVAVTAAQSTIEEQA
Chlorocebus aethiops   MSSSSWLLLSLVAVTAAQSTIEEQA
Macaca mulatta         MSGSSWLLLSLVAVTAAQSTIEEQA
Rhinolophus pearsonii  MSGSFWFLLSLVAVTAAQSTTEDRA
Rhinolophus sinicus    MSGSFWLLLSLVAVTTAQSTTEDRA
Paguma larvata         MSGSFWLLLSFAALTAAQSTTEELA
Felis catus            MSGSFWLLLSFAALTAAQSTTEELA
Mustela putorius furo  MLGSSWLLLSLAALTAAQSTTEDLA
Procyon lotor          MLGSSWLLLSLAALTAAQSTTEDLA
Nyctereutes procyonoides MSGSSWLLLSLAALTAAQS-TEDLV
```

**Figure 1.** Multiple sequence alignment of 11 mammalian ACE2 proteins. Only 25 amino acid sites from the N-terminus are shown, taken from Wei et al. [14].

The second case of suboptimal alignment is caused not by the progressive MSA, but by codon sequence alignment which takes one of two approaches. One approach is to translate codon sequences into amino acid sequences, align the amino acid sequences, and then map the codons to aligned amino acid sequences ([15,16], pp. 38–39). The other approach is to align codon sequences directly with a 64 × 64 scoring matrix, as is implemented in the PhyPA function of DAMBE [17,18]. A simple illustration of a suboptimal alignment obtained with the codon-based alignment is shown with three codon sequences (Figure 2A). Alignment 1 (Figure 2B) is obtained from codon-based alignment. It contains one triplet deletion and an A ↔ G substitution. In contrast, Alignment 2 (Figure 2C) contains only a triplet deletion and consequently represents a simper hypothesis with a higher alignment score than Alignment 1. Alignment 2 can be obtained from Alignment 1 in post-alignment adjustment.

Such suboptimal alignments (Figures 1 and 2) are not further refined by MSA programs such as the popular MUSCLE [6] and MAFFT [7]. However, one may formulate criteria to evaluate such suboptimal sites and make adjustment after the alignment. I present three methods based on three different criteria for this purpose: (1) sum-of-pairs score, (2) pairwise alignment inconsistency index, and (3) position weight matrix differential. The first two criteria are in general concordant, but they can conflict with the last criterion. However, the last one can often generate better MSA, leading to phylogenetic trees of higher likelihood than the other two criteria.

```
(A)      3   6   9  12  15
   S1 ATG CCC GTA TAA
   S2 ATG CCC GTG TCA TAA
   S3 ATG CCC GTG TCA TAG

(B) Alignment 1
   S1 ATG CCC GTA --- TAA
   S2 ATG CCC GTG TCA TAA
   S3 ATG CCC GTG TCA TAG
      *** *** **      ***

(C) Alignment 2
   S1 ATG CCC GT- --A TAA
   S2 ATG CCC GTG TCA TAA
   S3 ATG CCC GTG TCA TAG
      *** *** **    * ***
```

**Figure 2.** Suboptimal alignment of codon sequences. (**A**) Two unaligned codon sequences. (**B**) Alignment from codon-based alignment methods. (**C**) A better alignment based on alignment scores.

## 2. Criteria and Methods Used to Identify Suboptimal Sites in Alignments

Ideally, one would use maximum likelihood (ML) as a criterion for choosing the best alignment. From the same set of sequences, alternative alignment algorithmsa and scoring schemes may generate alternative alignments ($MSA_1$, $MSA_2$, . . . , $MSA_n$). From each of MSAs, one may reconstruct a maximum likelihood tree ($T_1$, $T_2$, . . . , $T_n$), with associated tree log-likelihood ($lnL_1$, $lnL_2$, . . . , $lnL_n$). $MSA_i$ is the best if $lnL_i$ is the largest. This application of the ML criterion needs to be conditional on the number of gaps in each alignment because an alignment with many additional indels to minimize nucleotide or amino acid mismatches will tend to increase likelihood. However, the real difficulty with integrating both an MSA and a phylogeny in a ML criterion is that it would be too slow to be practical [19,20].

I present three practical criteria and associated approaches for post-alignment adjustment. The first two has been criticized for not making use of information in all sequences simultaneous or not considering the evolutionary history among the sequences [13,21]. The last does use information from all sequences simultaneously, although it still does not make use of the evolutionary history of the sequences. I hope that the approaches presented here will foster more innovative approaches.

### *2.1. Sum-of-Pairs Score (SPS)*

Sum-of-pairs score (SPS) [22–26] has frequently been used as a criterion for evaluating alternative multiple alignment because of its conceptual simplicity. Each multiple alignment of $N$ sequences implies $N(N-1)/2$ pairwise alignments. SPS is simply the summation of all pairwise alignment scores without penalizing shared gaps. Obtaining SPS from MSAs is easy. All we need is a scoring scheme, i.e., gap-open and gap-extension penalties plus a match/mismatch matrix. A slight variation of SPS is the weighted SPS [13,21] in which alignment scores for some sequence pairs are weighted more heavily than others. This is expressed as

$$WSPS = \sum W_{ij} S_{ij} \quad (1)$$

which is reduced to SPS when $W_{ij} = 1$. MAFFT [7] uses multiple iterations of MSA refinement based on *WSPS* when either G-INS-i or L-INS-i option is chosen.

We need to evaluate Alignment 1 with "-T" (Figure 1) and Alignment 2 with "T-", occupying sites 20 and 21 in the *N. procyonoides* sequence. We only need to compute SPS for these two sites. Suppose we use BLOSUM62 score matrix and a gap penalty of −6 as our scoring scheme. In this particular case, we only need to compute pairwise alignment

scores between *N. procyonoides* and the other 10 species because all other pairwise scores are identical between the two alternative alignments (represented by constant C in Table 1).

**Table 1.** Sum-of-pairs scores for Alignment 1 (Figure 1) and an alternative Alignment 2 with "T-" occupying sites 20 and 21 in *N. procyonoides.* Only sites 20 and 21 in Figure 1 are considered.

| | T/-(1) | T/T(1) | T/I(1) | I/-(1) | SPS |
|---|---|---|---|---|---|
| Score(2) | −6 | 5 | −1 | −6 | |
| Alignment 1 | 10 | 6 | 4 | | −34 + C(3) |
| Alignment 2 | 6 | 10 | | 4 | −10 + C(3) |

(1) Amino acid pairs relevant for the calculation of SPS, (2) Gap penalty is −6, T/T match and T/I mismatch scores are 5 and −1, respectively, (3) C is a constant represents sum of pairwise scores from all sequences other than *N. procyonoides*.

For the 10 pairwise comparisons between *N. procyonoides* and the other 10 sequences at amino acid sites 20 and 21, Alignment 1 has 10 "T/-" pairs, 6 "T/T" pairs and 4 T/I pairs, yielding an SPS of −34 + C (Table 1). In contrast, Alignment 2 has 6 "T/-" pairs, 10 "T/T" pairs and 4 I/- pairs, yielding an SPS of −10 + C (Table 1). Therefore, Alignment 2 is better than Alignment 1.

The same approach can be applied to evaluate sites 9–12 in the two alternative alignments in Figure 2. Alignment 1 has both a triplet deletion and a nucleotide substitution, in contrast to Alignment 2 with only a triplet deletion but no nucleotide substitution, so SPS is greater for Alignment 2 than for Alignment 1.

### 2.2. Pairwise Alignment Inconsistency Index (PAI)

$N$ sequences have $N(N-1)/2$ pairwise alignments. Pairwise alignments can be inconsistent with each other and with those implied by MSA, which had been used to refine MSA before [17,27–31]. Designating $S_{ij}$ as the pairwise alignment score between sequences $i$ and $j$, and $S_{ij.MSA}$ as the equivalent score for paired alignment implied by the MSA, *PAI* is

$$PAI = \sum S_{ij} - \sum S_{ij.MSA} \quad (2)$$

Because $S_{ij}$ is from dynamic programming and consequently has the highest possible alignment score, whereas $S_{ij.MSA}$ is from the pairwise alignment implied by the progressive MSA, $S_{ij} \geq S_{ij.MSA}$. A poor MSA will have a larger *PAI* than a good MSA. For the 11 amino acid sequences (Figure 1) with Alignment 1 and Alignment 2 as defined before, we only need to compare the pairwise alignments between *N. procyonoides* and the other 10 sequences for sites 20 and 21. *PAI* for Alignment 1 is greater than *PAI* for Alignment 2 by a difference of 24. We conclude that Alignment 1 is worse than Alignment 2. The advantage of using *PAI* is that all $S_{ij}$ values are already computed in first guide tree during MSA, so there is little computational overhead.

### 2.3. Position Weight Matrix Differential (PWMD)

Position weight matrix (PWM) [32–36] was originally introduced into biology for characterizing regulatory motifs as components of regulons [37,38]. Its computation, as well as associated significance tests, has previously been illustrated numerically in great detail [37,39]. PWM scores (*PWMSs*) have been suggested as possible metrics for evaluating alternative MSAs [37]. A PWM can be generated from an MSA, and *PWMS* can be computed for each sequence. When one or more nucleotides or amino acids are shifted along indels, the difference in *PWMS* before and after the shifting is

$$PWMD = \sum PWMS_{after} - \sum PWMS_{before} \quad (3)$$

Any nucleotide or amino acid shift that results in a positive *PWMD* is desirable. The 11 ACE2 sequences in Figure 1 have an alignment length of 805, so the resulting PWM is a 20 × 805 matrix. However, we only need to look at sites 20 and 21 (Table 2). Site 20

is occupied by amino acid T, so only amino acid T has a positive value at site 20. Site 21 is occupied by both T and I, so only these two amino acids have positive values at site 21 (Table 2). The PWM from either Alignment 1 or Alignment 2 suggest that we should put amino acid T at site 20 instead of at site 21 in *N. procyonoides.* For the PWM derived from Alignment 1, placing T at site 20 instead of 21 yields a PWMD of 0.6481 (=4.2457−3.5976, Table 2).

**Table 2.** Partial position weight matrix for 11 aligned ACE2 sequences, generated from DAMBE [40] using default options for pseudocounts and background frequencies. Only sites 20 and 21 are included.

| | **Alignment 1** | | **Alignment 2** | |
|---|---|---|---|---|
| **AA** | **Site 20** | **Site 21** | **Site 20** | **Site21** |
| A | −3.4621 | −3.4621 | −3.4621 | −3.4621 |
| R | −3.4632 | −3.4632 | −3.4632 | −3.4632 |
| N | −3.4620 | −3.4620 | −3.4620 | −3.4620 |
| D | −3.4625 | −3.4625 | −3.4625 | −3.4625 |
| C | −3.4757 | −3.4757 | −3.4757 | −3.4757 |
| Q | −3.4632 | −3.4632 | −3.4632 | −3.4632 |
| E | −3.4616 | −3.4616 | −3.4616 | −3.4616 |
| G | −3.4625 | −3.4625 | −3.4625 | −3.4625 |
| H | −3.4673 | −3.4673 | −3.4673 | −3.4673 |
| I | −3.4628 | 2.9353 | −3.4628 | 2.9353 |
| L | −3.4612 | −3.4612 | −3.4612 | −3.4612 |
| K | −3.4624 | −3.4624 | −3.4624 | −3.4624 |
| M | −3.4645 | −3.4645 | −3.4645 | −3.4645 |
| F | −3.4629 | −3.4629 | −3.4629 | −3.4629 |
| P | −3.4628 | −3.4628 | −3.4628 | −3.4628 |
| S | −3.4619 | −3.4619 | −3.4619 | −3.4619 |
| T | 4.1089 | 3.5976 | 4.2457 | 3.3770 |
| W | −3.4649 | −3.4649 | −3.4649 | −3.4649 |
| Y | −3.4632 | −3.4632 | −3.4632 | −3.4632 |
| V | −3.4621 | −3.4621 | −3.4621 | −3.4621 |

The previous presentation of the three approaches might mislead the reader to think that the two criteria are all consistent with each other. This is unfortunately not the case. While the first two approaches are generally consistent with each other, they often conflict with the third criterion (PWMD). I will illustrate this with a more realistic data set with alignment of huntingtin (HTT) proteins.

## 3. A comparison of Methods with Huntingtin Sequence Alignment

Huntington's disease is associated with the length of glutamine (Q, encoded by CAG and CAA codons) repeats in the huntingtin (HTT) protein. The expansion and shrinking of $(CAG)_n$, where the subscript $n$ is the number of consecutive CAG codons, is caused by strand slippage during DNA replication [41]. Huntington's disease typically manifests with $n > 37$. The longer the repeats, the earlier the disease onset [42].

I downloaded 20 primate HTT protein sequences and aligned them using MAFFT [7] with the slow but accurate G-INS-i option that uses progressive alignment with multiple iterative refinements based on weighted sum-of-pairs score as defined in Equation (1). The aligned sequences are included in FASTA format in the Supplemental file Primate_HTT_MAFFT.fas.zip. The MSA contains 3156 aligned sites, but only the first 53 sites are shown in Figure 3A for illustrating the PWM-based post-alignment refinement. This alignment is contrasted with an alternative alignment (Figure 3B), based on PWMD that I will explain later.

```
XP_016806693_Pan_troglodytes      (A)  MATLEKLMKAFESLKSF----------QQQQQQQQQQQ-----QPPPPPPPLP
XP_003813027_Pan_paniscus              MATLEKLMKAFESLKSF----------QQQQQQQQQQ-------PPPPPPPLP
NP_001375421_Homo_sapiens              MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQ-----PPPPPPPPP
XP_030865871_Gorilla_gorilla_gorilla   MATLEKLMKAFESLKSF----------QQQQQQQQQQQ-----PPPPPPPLP
XP_024101641_Pongo_abelii              MATLEKLMKAFESLKSF----------QQQQQQQQQQQP-----PPPPPPPPP
XP_030657254_Nomascus_leucogenys       MATLEKLMKAFESLKSF----------QQQQQQQQQQ-------PPPPPPPPP
XP_031991928_Hylobates_moloch          MATLEKLMKAFESLKSF----------QQQQQQQQQQ-------PPPPPPPPP
XP_028704080_Macaca_mulatta            MATLEKLMKAFESLKSF---------QQQQQQQQQQQP-----PPPPPPPPP
XP_011743216_Macaca_nemestrina         MATLEKLMKAFESLKSF----------QQQQQQQQQQQ-----PPPPPPPPP
XP_011914562_Cercocebus_atys           MATLEKLMKAFESLKSF----------QQQQQQQQQQQP-----PPPPPPPPP
XP_021793897_Papio_anubis              MATLEKLMKAFESLKSF-------QQQQQQQQQQQQP-----PPPPPPPPP
XP_025241155_Theropithecus_gelada      MATLEKLMKAFESLKSF-------QQQQQQQQQQQQP-----PPPPPPPPP
XP_033069457_Trachypithecus_francoisi  MATLEKLMKAFESLKSF----------QQQQQQQQQQQ-----QPPPPPPPP
XP_030775944_Rhinopithecus_roxellana   MATLEKLMKAFESLKSF---------QQQQQQQQQQQP-----PPPPPPPPP
XP_023062590_Piliocolobus_tephrosceles MATLEKLMKAFESLKSF----------QQQQQQQQQQP-----PPPPPPPPP
XP_032109912_Sapajus_apella            MATLEKLMKAFESLKSF----------QQQQQQQQQQQ---PPPPPPPPPPP
XP_039324056_Saimiri_boliviensis       MATLEKLMKAFESLKSF----------QQQQQQQQQQQQPPPPPPPPPPPPP
XP_012322257_Aotus_nancymaae           MATLEKLMKAFESLKSF----------QQQQQQQQQQQ-----PPPPPPPPP
XP_012592808_Microcebus_murinus        MATLEKLMKAFESLKSF----------QQQQQQQ-----------PPPPPPP
XP_012510503_Propithecus_coquereli     MATLEKLMKAFESLKSF----------QQQQQQ------------PPPPPPQP
                                       12345678901234567890123456789012345678901234567890123
XP_016806693_Pan_troglodytes      (B)  MATLEKLMKAFESLKSF---------QQQQQQQQQQQ-----PPPPPPPLP
XP_003813027_Pan_paniscus              MATLEKLMKAFESLKSF-----------QQQQQQQQQQ-----PPPPPPPLP
NP_001375421_Homo_sapiens              MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQ----PPPPPPPPP
XP_030865871_Gorilla_gorilla_gorilla   MATLEKLMKAFESLKSF----------QQQQQQQQQQQ----PPPPPPPLP
XP_024101641_Pongo_abelii              MATLEKLMKAFESLKSF-----------QQQQQQQQQQ---PPPPPPPPP
XP_030657254_Nomascus_leucogenys       MATLEKLMKAFESLKSF-----------QQQQQQQQQQ-----PPPPPPPP
XP_031991928_Hylobates_moloch          MATLEKLMKAFESLKSF-----------QQQQQQQQQQ-----PPPPPPPP
XP_028704080_Macaca_mulatta            MATLEKLMKAFESLKSF----------QQQQQQQQQQQ---PPPPPPPPP
XP_011743216_Macaca_nemestrina         MATLEKLMKAFESLKSF----------QQQQQQQQQQ----PPPPPPPPP
XP_011914562_Cercocebus_atys           MATLEKLMKAFESLKSF-----------QQQQQQQQQQ---PPPPPPPPP
XP_021793897_Papio_anubis              MATLEKLMKAFESLKSF--------QQQQQQQQQQQQ---PPPPPPPPP
XP_025241155_Theropithecus_gelada      MATLEKLMKAFESLKSF--------QQQQQQQQQQQQ---PPPPPPPPP
XP_033069457_Trachypithecus_francoisi  MATLEKLMKAFESLKSF----------QQQQQQQQQQ-----PPPPPPPP
XP_030775944_Rhinopithecus_roxellana   MATLEKLMKAFESLKSF----------QQQQQQQQQQ---PPPPPPPPP
XP_023062590_Piliocolobus_tephrosceles MATLEKLMKAFESLKSF-----------QQQQQQQQQQ---PPPPPPPPP
XP_032109912_Sapajus_apella            MATLEKLMKAFESLKSF----------QQQQQQQQQQ--PPPPPPPPPP
XP_039324056_Saimiri_boliviensis       MATLEKLMKAFESLKSF----------QQQQQQQQQQPPPPPPPPPPP
XP_012322257_Aotus_nancymaae           MATLEKLMKAFESLKSF----------QQQQQQQQQQ----PPPPPPPPP
XP_012592808_Microcebus_murinus        MATLEKLMKAFESLKSF--------------QQQQQQ------PPPPPPP
XP_012510503_Propithecus_coquereli     MATLEKLMKAFESLKSF---------------QQQQQQ------PPPPPPQP
```

**Figure 3.** N-terminus of 20 aligned HTT sequences, with the site numbering in the middle. (**A**) Alignment from MAFFT [7] with optimized options. (**B**) One of the alternative alignments refined with the PWMD criterion.

Both SPS and PAI indices would favor the alignment in Figure 3A against that in Figure 3B. For example, if we use a gap open (GO) penalty of 20, a gap extension penalty of 2, and the BLOSUM62 score matrix, then SPS is 31,007 for the MSA in Figure 3A, but only 29,178 for the MSA in Figure 3B. This is clearly seen from the pairwise alignment between the first (*Pan troglodytes*) and the third (*Homo sapiens*) sequences. There is only one GO in the alignment between these two sequences in Figure 3A but two GOs in Figure 3B. A number of similar differences contribute to a much larger SPS for the alignment in Figure 3A than that in Figure 3B. Therefore, the SPS and PAI indices would favor the alignment in Figure 3A against the alignment in Figure 3B.

A likelihood method, however, would favor the alignment in Figure 3B against that in Figure 3A. We may reconstruct a phylogenetic tree from each of these two alignments using PhyML [43] with (1) the LG substitution matrix and (2) a constant rate of amino acid substitution over sites. This yields a tree log-likelihood (lnL) of −126.69 for the alignment in Figure 3A but −106.77 for the alignment in Figure 3B. One may change substitution matrices but the tree lnL is consistently greater for the alignment in Figure 3B than for the alignment in Figure 3A (Table 3).

**Table 3.** Tree log-likelihood values for the two multiple sequence alignments in Figure 3, obtained with PhyML and three different substitution matrices.

| | **Substitution Matrix** | | |
|---|---|---|---|
| **Alignment** | **LG** | **JTT** | **BLOSUM62** |
| in Figure 3A | −126.6903 | −122.6004 | −126.7423 |
| in Figure 3B | −106.7703 | −105.2280 | −106.9387 |

One might argue that the ML criterion in Table 3 is not fair because there are more amino acid substitutions in the alignment in Figure 3A than in Figure 3B. This is a valid criticism. However, one may defend the alignment in Figure 3B in two ways. First, the alignment in Figure 3B did not add indels to reduce amino acid substitutions. In fact, the alignment in Figure 3B has 20 fewer gaps than that in Figure 3A. Second, the alignment in Figure 3B suggests that the expansion/shrinkage of repeated amino acid Q occurs more frequently than amino acid replacement. This is consistent with the documented strand slippage during DNA replication in generating length variations of $(CAG)_n$ tracts in the *HTT* gene [41].

The alignment in Figure 3B is one of the optimal alignments based on the PWMD criterion. It highlights a case where the criterion of PWMD conflicts with SPS and PAI. I outline below the steps involved in post-alignment adjustment involving PWMD.

**Step 1:** From the alignment of HTT sequences obtained from MAFFT with 3156 aligned sites, one can compute the 20 × 3156 PWM. Part of the PWM, with the relevant sites and the two relevant amino acids (Q and P) is shown in Table 4. The PWM values in the "Q" column (Table 4) state that $Q_6$ (where the subscript 6 is the number of consecutive Qs in a sequence) should be placed at sites 28–33 (where the PWM values are the largest), $Q_7$ at sites 28–34, $Q_{10}$ at sites 28–37, $Q_{11}$ at sites 28–38, $Q_{21}$ at sites 18–38, and so on.

**Table 4.** Part of the 20 × 3156 position weight matrix obtained with default options for pseudocounts and background frequencies in DAMBE [40]. Only sites 18 to 44 from 3156 aligned sites are shown, with only two amino acids (Q and P) out of 20. Site numbers are as in the alignment in Figure 3A.

| Site | Q | P |
|---|---|---|
| 18 | −0.0374 | −4.3223 |
| 19 | −0.0374 | −4.3223 |
| 20 | −0.0374 | −4.3223 |
| 21 | −0.0374 | −4.3223 |
| 22 | −0.0374 | −4.3223 |
| 23 | −0.0374 | −4.3223 |
| 24 | −0.0374 | −4.3223 |
| 25 | 1.4974 | −4.3223 |
| 26 | 1.4974 | −4.3223 |
| 27 | 2.2241 | −4.3223 |
| 28 | 4.2125 | −4.3223 |
| 29 | 4.2125 | −4.3223 |
| 30 | 4.2125 | −4.3223 |
| 31 | 4.2125 | −4.3223 |
| 32 | 4.2125 | −4.3223 |
| 33 | 4.2125 | −4.3223 |
| 34 | 4.1387 | −4.3223 |
| 35 | 4.0609 | −4.3223 |
| 36 | 4.0609 | −4.3223 |
| 37 | 4.0609 | −4.3223 |
| 38 | 2.8964 | 2.5727 |
| 39 | −0.0374 | −4.3223 |
| 40 | −4.3224 | −0.1637 |
| 41 | −4.3224 | −0.1637 |
| 42 | −4.3224 | 0.7953 |
| 43 | −4.3224 | 0.7953 |
| 44 | 0.9251 | 3.4601 |

The Step 1 refinements result in a favorable PWMD of 9.198. The alignment after Step 1 is shown in Figure 4A.

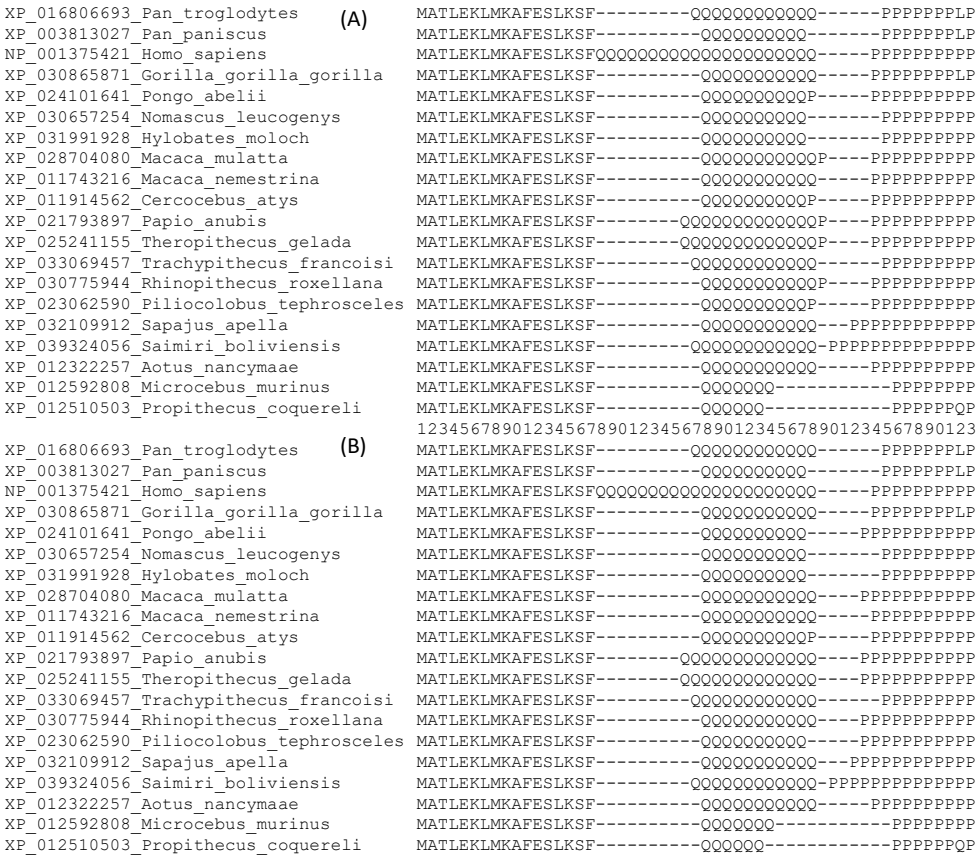


**Figure 4.** Illustration of PWM-based refinement of sequence alignment based on the N-terminus of 20 aligned HTT sequences, with the site numbering in the middle. (**A**) Alignment after Step 1 refinement. (**B**) Alignment after Step 2 refinement, except that the shared gap at site 39 has not yet been deleted.

**Step 2:** The alignment after Step 1 (Figure 4A) has three P residues at site 38 mixed with 12 Q residues. At site 43, there are two P residues without any other amino acids. Moving these three P residues from site 38 in Figure 4A to site 43 increases the PWMD. Similarly, shifting the four P residues from site 39 in Figure 4A to site 43 also increases the PWMD. These refinements result in the alignment in Figure 4B. Such refinements also result in a shared gap at site 39 (Figure 4B) which can be deleted. These refinements yield a further gain of PWMD of 101.4509 (relative to the alignment in Figure 4A).

After Step 2, no further refinement will result in a positive PWMD. Note that the alignment in Figure 4B, after deleting the shared gap at site 39, looks different from that in Figure 3B, but they are equally good based on the PWMD criterion (i.e., changing one to the other will have PWMD = 0). In fact, there are many alternative alignments that are equally good to the alignment in Figure 3B based on the PWMD criterion. They also produce the same tree lnL.

## 4. Discussion

While the PWMD criterion appears promising for post-alignment adjustment, this paper is no more than a proof of concept. There are obviously cases more complicated than the two illustrative examples. Such cases may require multiple iterations of PWM computation and MSA refinement. However, the PWM-based approach does feature three advantages. Firstly, it is conceptually simple. Secondly, it uses information from all sequences. Thirdly, it is fast because PWM requires little computation time, so multiple iterations of refinements can be accomplished in little time.

As I have illustrated, the PWMD criterion can conflict with the SPS and PAI criteria. This is disconcerting given that SPS [22–26], its weighted form [13,21], as well as PAI [17,27–31] have been used frequently both in generating MSAs and in iterative MSA refinement. One may argue that PWMD uses information from all sequences, so it is preferable over the SPS and PAI criteria which are based on information from pairwise alignment and have been criticized in this context [13,21]. It might indeed be time to reconsider SPS and PAI as criteria for MSA refinement.

None of the three criteria illustrated here incorporate the evolutionary history of the aligned sequences. Unfortunately, including an inference of evolutionary history among the sequences in the MSA refining process will invariably demand intensive computation [19,20]. The PWMD criterion, although making use of information from all sequences, implicitly treats all sequences equally as if they were from a star tree. Whether this feature of PWMD might have the benefit of not biasing subsequent phylogenetic estimation would require further studies.

The PWMD criterion has not yet been implemented in any publicly available software for post-alignment adjustment, so its performance has not been explored in any significant scale. Whether it will be adopted by the research community depends on not only the theoretical justification, but also the implementation of the method in user-friendly software packages.

I should emphasize the effect of taxon sampling on sequence alignment and post-alignment adjustment, as this effect is important but often neglected. Take the alignment in Figure 1, for example. The ACE2 sequence in *Mus musculus*, which is not in the alignment, is "LT" at sites 20 and 21. If I remove the first four (primate) ACE2 sequences in Figure 1 and add many ACE2 sequences similar to that of *M. musculus*, then the three post-alignment adjustment approaches would all favor "-T" at sites 20 and 21 against the alternative "T-" in *N. procyonoides*, contrary to the post-alignment adjustment that we have made before. This again highlights the need to incorporate evolutionary history in post-alignment adjustments. If *M. musculus* is phylogenetically closer to *N. procyonoides* than the first four primate species in Figure 1, then we should keep "-T" at sites 20 and 21 in *N. procyonoides.* In contrast, if the first four primate species are phylogenetically closer to *N. procyonoides*, then we should revise "-T" in *N. procyonoides* (Figure 1) to "T-". While many researchers have highlighted the effect of taxon sampling on phylogenetic reconstruction [44–46], few have so far recognized the fact that the effect of taxon sampling is often seeded in multiple-sequence alignment.

## 5. Conclusions

I illustrated the importance of post-alignment adjustment, outlined criteria to rapidly evaluate alternative alignments, and presented three approaches towards post-alignment adjustment. I highlighted the potential of the position weight matrix approach and illustrated its applications for refining several sets of real sequences. The problem of post-alignment adjustment is not fully solved. I hope that my presentation of the problem and the directions towards a solution will stimulate further research in this rapidly developing field.

**Supplementary Materials:** The following are available online at https://www.mdpi.com/article/10.3390/genes12111809/s1, Primate_HTT_MAFFT.fas.

**Funding:** This research was funded by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC, RGPIN/2018-03878) of Canada. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** I thank Q. Yang, L. Li and members of Xia Lab for comments and discussion. Two reviewers provided constructive feedback. I am particularly grateful to one of the reviewers who patiently corrected my writing in numerous places. I wish I could serve equally well as a reviewer.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

1. Blackburne, B.P.; Whelan, S. Class of multiple sequence alignment algorithm affects genomic analysis. *Mol. Biol. Evol.* **2013**, *30*, 642–653. [CrossRef] [PubMed]
2. Kumar, S.; Filipski, A. Multiple sequence alignment: In pursuit of homologous DNA positions. *Genome Res.* **2007**, *17*, 127–135. [CrossRef] [PubMed]
3. Wong, K.M.; Suchard, M.A.; Huelsenbeck, J.P. Alignment uncertainty and genomic analysis. *Science* **2008**, *319*, 473–476. [CrossRef] [PubMed]
4. Noah, K.; Hao, J.; Li, Y.; Sun, X.; Foley, B.T.; Yang, Q.; Xia, X. Major revisions in arthropod phylogeny through improved supermatrix, with support for two possible waves of land invasion by chelicerates. *Evol. Bioinform.* **2020**, *16*, 1176934320903735. [CrossRef]
5. Xia, X. *A Mathematical Primer of Molecular Phylogenetics*; CRC Press: New York, NY, USA, 2020; p. 380.
6. Edgar, R.C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **2004**, *5*, 113. [CrossRef] [PubMed]
7. Katoh, K.; Asimenos, G.; Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* **2009**, *537*, 39–64.
8. Hogeweg, P.; Hesper, B. The alignment of sets of sequences and the construction of phylogenetic trees: An integrated method. *J. Mol. Evol.* **1984**, *20*, 175–186. [CrossRef]
9. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673–4680. [CrossRef] [PubMed]
10. Xia, X. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. *Mol. Biol. Evol.* **2020**, *37*, 2699–2705. [CrossRef]
11. Xia, X. Dating the Common Ancestor from an NCBI Tree of 83688 High-Quality and Full-Length SARS-CoV-2 Genomes. *Viruses* **2021**, *13*, 1790. [CrossRef] [PubMed]
12. Xia, X. Sequence Alignment. In *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics*; Springer: Cham, Switzerland, 2018; pp. 33–75.
13. Higgins, D.; Lemey, P. Multiple sequence alignment. In *The Phylogenetic Handbook*; Lemey, P., Salemi, M., Vandamme, A.M., Eds.; Cambridge University Press: Cambridge, UK, 2009; pp. 68–108.
14. Wei, Y.; Aris, P.; Farookhi, H.; Xia, X. Predicting mammalian species at risk of being infected by SARS-CoV-2 from an ACE2 perspective. *Sci. Rep.* **2021**, *11*, 1702. [CrossRef] [PubMed]
15. Xia, X. *Data Analysis in Molecular Biology and Evolution*; Kluwer Academic Publishers: Boston, UK, 2000; p. 277.
16. Xia, X.; Xie, Z. DAMBE: Software package for data analysis in molecular biology and evolution. *J. Hered.* **2001**, *92*, 371–373. [CrossRef] [PubMed]
17. Xia, X. PhyPA: Phylogenetic method with pairwise sequence alignment outperforms likelihood methods in phylogenetics involving highly diverged sequences. *Mol. Phylogenet. Evol.* **2016**, *102*, 331–343. [CrossRef] [PubMed]
18. Xia, X. DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. *J. Hered.* **2017**, *108*, 431–437. [CrossRef] [PubMed]
19. Sankoff, D.; Cedergren, R.J.; Lapalme, G. Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. *J. Mol. Evol.* **1976**, *7*, 133–149. [CrossRef] [PubMed]
20. Vingron, M.; von Haeseler, A. Towards integration of multiple alignment and phylogenetic tree construction. *J. Comput. Biol.* **1997**, *4*, 23–34. [CrossRef] [PubMed]
21. Edgar, R.C.; Batzoglou, S. Multiple sequence alignment. *Curr. Opin. Struct. Biol.* **2006**, *16*, 368–373. [CrossRef] [PubMed]
22. Althaus, E.; Caprara, A.; Lenhof, H.P.; Reinert, K. Multiple sequence alignment with arbitrary gap costs: Computing an optimal solution using polyhedral combinatorics. *Bioinformatics* **2002**, *18*, S4–S16. [CrossRef]
23. Reinert, K.; Stoye, J.; Will, T. An iterative method for faster sum-of-pairs multiple sequence alignment. *Bioinformatics* **2000**, *16*, 808–814. [CrossRef]
24. Stoye, J.; Moulton, V.; Dress, A.W. DCA: An efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. *Comput. Appl. Biosci.* **1997**, *13*, 625–626. [CrossRef]
25. Lipman, D.J.; Altschul, S.F.; Kececioglu, J.D. A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 4412–4415. [CrossRef] [PubMed]
26. Gupta, S.K.; Kececioglu, J.D.; Schaffer, A.A. Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *J. Comput. Biol.* **1995**, *2*, 459–472. [CrossRef] [PubMed]
27. Notredame, C.; Higgins, D.G.; Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **2000**, *302*, 205–217. [CrossRef] [PubMed]

28. Floden, E.W.; Tommaso, P.D.; Chatzou, M.; Magis, C.; Notredame, C.; Chang, J.M. PSI/TM-Coffee: A web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. *Nucleic Acids Res.* **2016**, *44*, W339–W343. [CrossRef] [PubMed]
29. Magis, C.; Taly, J.F.; Bussotti, G.; Chang, J.M.; Di Tommaso, P.; Erb, I.; Espinosa-Carrasco, J.; Notredame, C. T-Coffee: Tree-based consistency objective function for alignment evaluation. *Methods Mol. Biol.* **2014**, *1079*, 117–129.
30. Chang, J.M.; Di Tommaso, P.; Notredame, C. TCS: A new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol. Biol. Evol.* **2014**, *31*, 1625–1637. [CrossRef]
31. Gotoh, O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.* **1996**, *264*, 823–838. [CrossRef]
32. Staden, R. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* **1984**, *12*, 505–519. [CrossRef]
33. Stormo, G.D.; Schneider, T.D.; Gold, L. Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res.* **1986**, *14*, 6661–6679. [CrossRef]
34. Hertz, G.Z.; Hartzell, G.W., III; Stormo, G.D. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* **1990**, *6*, 81–92. [CrossRef]
35. Claverie, J.M.; Audic, S. The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.* **1996**, *12*, 431–439. [CrossRef] [PubMed]
36. Hertz, G.Z.; Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **1999**, *15*, 563–577. [CrossRef] [PubMed]
37. Xia, X. Position weight matrix and Perceptron. In *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics*; Springer: Cham, Switzerland, 2018; pp. 77–98.
38. Xia, X. Beyond Trees: Regulons and Regulatory Motif Characterization. *Genes* **2020**, *11*, 995. [CrossRef]
39. Xia, X. Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction. *Scientifica* **2012**, *2012*, 917540. [CrossRef] [PubMed]
40. Xia, X. DAMBE7: New and improved tools for data analysis in molecular biology and evolution. *Mol. Biol. Evol.* **2018**, *35*, 1550–1552. [CrossRef] [PubMed]
41. Xu, P.; Pan, F.; Roland, C.; Sagui, C.; Weninger, K. Dynamics of strand slippage in DNA hairpins formed by CAG repeats: Roles of sequence parity and trinucleotide interrupts. *Nucleic Acids Res.* **2020**, *48*, 2232–2245. [CrossRef]
42. Wexler, N.S.; Lorimer, J.; Porter, J.; Gomez, F.; Moskwitz, C.; Shackell, E.; Karen, M.; Penchaszadeh, G.; Roberts, S.A.; Gayan, J.; et al. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 3498–3503.
43. Guindon, S.; Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **2003**, *52*, 696–704. [CrossRef] [PubMed]
44. Heath, T.A.; Zwickl, D.J.; Kim, J.; Hillis, D.M. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. *Syst. Biol.* **2008**, *57*, 160–166. [CrossRef]
45. Poe, S.; Swofford, D.L. Taxon sampling revisited. *Nature* **1999**, *398*, 299–300. [CrossRef]
46. Zwickl, D.J.; Hillis, D.M. Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* **2002**, *51*, 588–598. [CrossRef] [PubMed]