On alignment shift and its measures

Melissa Cline
Kevin Karplus

UCSC-CRL-97-27
February 26, 1998

Baskin Center for
Computer Engineering & Information Sciences
University of California, Santa Cruz
Santa Cruz, CA 95064 USA

ABSTRACT

This document proposes a new measure for comparing two alignments, referred to as the *shift score.* This score has a range of $-\epsilon$ to 1, where ϵ is a small number used as a parameter to the scoring algorithm. A score of 1 is attained only if the two alignments are identical. The shift score is symmetric with respect to the alignments—no distinguished sequence or alignment is needed. The score incorporates both coverage measures and shift error measures into a single number. The shift score is easily computed, and one can find the optimum subalignment of a candidate alignment (the subalignment that maximizes the shift score) with a simple greedy algorithm.

1. Introduction

This document discusses our measurement of alignment shift, and two measures that can be derived from alignment shift: mean shift error and shift score. Section 2 presents a precise definition of shift, including a description of when it can and cannot be measured.

In some candidate alignment, shift is a measure of the distance between two pairs of residues aligned in a "correct" pairwise alignment. The distance is in separation along the protein backbone, as illustrated in Figure 2.1.

Shift may be positive or negative, depending on the relative positions in the candidate alignment of the two residues. If a residue is aligned to the left of its correct position, its shift is negative. If it is aligned to the right of its correct position, its shift is positive.

To summarize the overall amount of shifting in an alignment, individual shift measures are combined into an overall shift score. Section 3 describes two shift scores: mean shift error and shift score. The mean shift error, detailed in Section 3.1, is the most common statistic on overall shift [Bry96, MBB97]. Section 3.2 presents the shift score measure, and makes a case on why it is more useful than mean shift.

2. Definition of shift

2.1 Terminology

Shift is inherently a measure of the differences between two alignments: a *reference alignment* and a *candidate alignment.* The reference alignment is considered the standard of correctness, while the candidate alignment is thought of as the alignment that might be in error. Further, shift is a measure of differences in pairwise alignments. The two sequences aligned are referred to as the *template sequence* and *target sequence.* In a structure-prediction scenario, the template sequence would have a known structure, and the target sequence would be a sequence of unknown structure aligned to the template as a prediction. In more general terms, the template sequence serves as the basis of the alignment, and the target sequence is aligned to the template in a manner that might not be correct.

2.2 Definition of the shift of a single residue

Consider a pair of residues A and B aligned in the reference alignment. If in the candidate alignment, residue B is aligned to some template residue C rather than A, then the shift of B is defined as the number of positions between C and A in the template sequence, as illustrated in Figure 2.1. Note that the shift can be positive or negative, depending on the direction of the shift. A positive shift moves residue B closer to the C-terminus of the sequences, or to the right in a standard visualization of the alignment. Note also that if residue B is not aligned in either the reference or candidate alignment, then the shift of B is undefined.

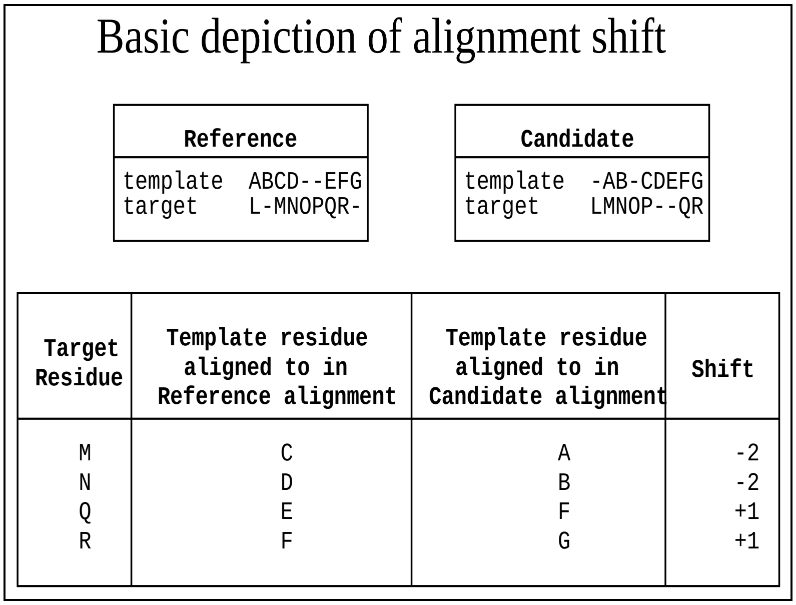
Shift is measured at each available position in the target sequence. If the target and template sequences are swapped, the shift measures can change radically due to insertions in one of the sequences. Similarly, if the reference and candidate alignments are swapped, shift measures will change in sign and can also change in magnitude. Thus, shift is sensitive to the assignment of target and template sequence, and the assignment of reference and candidate alignment.

Because shift is measured by sequence positions, rather than by alignment columns, the shift of a residue is not affected by the number of deletions between the residues to which it is aligned in the reference and candidate alignment. For instance, in Figure 2.1, target residue M has shifted by two template sequence positions and three alignment columns. Its shift is -2, indicating that it has moved two template sequence positions in the direction of the start of the sequence.

2.3 What about differing sequences?

In practice, we cannot assume that in the two different alignments, the two versions of the same sequence are the same. One alignment might reflect regions that have been resolved since the other alignment was built, and might contain additional residues.

If the two alignments contain two different versions of the same sequence, we cannot assume that one version is correct and the other is not. Instead, We address this matter by aligning the two versions of the sequence to each other. When a subsequence is missing



Reference:

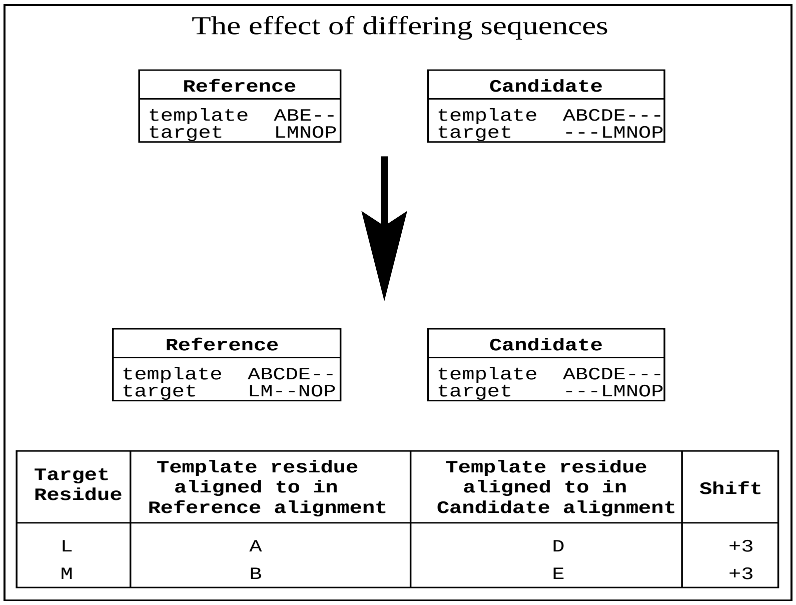
```
template  ABCD--EFG
target    L-MNOPQR-
```

Candidate:

```
template  -AB-CDEFG
target    LMNOP--QR
```

Target Residue	Template residue aligned to in Reference alignment	Template residue aligned to in Candidate alignment	Shift
M	C	A	-2
N	D	B	-2
Q	E	F	+1
R	F	G	+1

Figure 2.1: Basic illustration of the shift of a single residue. Shift is measured for target sequence residues aligned in both the reference and candidate alignment. It refers to the number of template sequence positions from where the residue is aligned in the reference alignment to where it's aligned in the candidate alignment.



Reference:

```
template  ABE--
target    LMNOP
```

Candidate:

```
template  ABCDE---
target    ---LMNOP
```

Reference:

```
template  ABCDE--
target    LM--NOP
```

Candidate:

```
template  ABCDE---
target    ---LMNOP
```

Target Residue	Template residue aligned to in Reference alignment	Template residue aligned to in Candidate alignment	Shift
L	A	D	+3
M	B	E	+3

Figure 2.2: When the reference and candidate alignments show two differing versions of the same alignment, the two versions are aligned to each other and all residues missing from one sequence are added to it. When shift measures are made, they can reflect these added residues.

from one version of the sequence, we add it to the other version as specified by this "same-sequence" alignment. This is illustrated in Figure 2.2.

When shift is measured, it is measured using these augmented alignments. The number of sequence positions shifted includes both residues present in and residues absent from the original aligned sequences.

3. Statistics on overall alignment shift

This section describes the ways in which shift measures can be combined into one measure of alignment quality. There are two scenarios in which a quantifier of alignment quality is vital. The first is a scenario such as a CASP contest, in which there are many different alignments of the same sequence and the best alignments must be chosen objectively. The second scenario is comparing methods for generating alignments; over a test set of hundreds of alignments, the goal is to determine which method produces the best alignments consistently. The first application requires only that alignments that are better from a biological standpoint get better shift scores. The second application requires that the score be uniformly interpretable over a wide range of alignment length and quality.

3.1 Mean shift

Mean shift is the most common statistic based on alignment shift [Bry96, MBB97]. Mean shift is simply the mean of the absolute shift at all positions in the target sequence for which shift can be measured. It is computed as follows:

$$\begin{aligned} N &= \text{Number of target sequence residues with shift data available} \\ \text{shift}_\text{i} &= \text{Shift value for shifted residue } i \\ \text{MSE} &= \text{The mean shift error for the candidate alignment} \\ &= \frac{1}{N}\sum_{i=1}^{N}|\text{shift}_\text{i}| \end{aligned}$$

While mean shift is a popular alignment quality measure, it has a few drawbacks. First, mean shift is sensitive to the selection of candidate versus reference alignment, and target versus template sequence. If any of these assignments are changed, mean shift can change dramatically. This behavior gives mean shift an arbitrary nature, since the measure only has meaning for a specific assignment of reference, candidate, template, and target. In situations where there is no alignment that can be considered the gold standard, this arbitrary nature is unfortunate.

Second, mean shift error cannot be taken alone as a measure of alignment quality. Mean shift error is closely related to *coverage*: for all target sequence residues aligning to template residues in the candidate alignment, coverage is the percentage that align to template residues in the reference alignment as well. If an alignment includes only the “easy” sections such as sections of high pairwise identity, it can have a low mean shift error and low coverage. If an alignment includes the more challenging regions, it can have high coverage and a high mean shift error. Thus, mean shift error is not a useful measure without coverage. Even when mean shift error and coverage are both reported, they have to be combined to produce a useful and objective measure of alignment quality, and there is no agreed-upon method for combining them.

On the third point, consider that the existence of one correct structural alignment between two proteins is debatable [God96]. When two proteins are aligned by different structural alignment tools, the resulting alignments often contain regions shifted by a few positions. Since there is not one "standard of truth", a penalty levied on small shifts might not be useful. Consider two different candidate alignments compared to one single reference alignment: one with a number of small shifts, and one with large regions of zero shift and small regions of a very large shift. Both alignments could be close in their mean shift error, though the first might arguably be the better alignment. The problem is that mean shift cannot ignore differences so small that they might not be significant.

3.2 Shift score

We are introducing a new measure which we believe overcomes the limitations of the mean shift score. The *shift score* is computed as follows:

$$
\begin{array}{rcl}
A, B & = & \text{The two sequences to be aligned} \\
|R| & = & \text{Number of pairs of aligned residues in the reference alignment} \\
|C| & = & \text{Number of pairs of aligned residues in the candidate alignment} \\
N_x & = & \text{Number of residues in sequence } x \\
i_x & = & \text{Residue number i in sequence } x \\
\mathrm{shift(i_x)} & = & \text{Shift of residue } i_x\text{, if defined} \\
s(i_x) & = & \left\{ \begin{array}{ll} \frac{1+\epsilon}{1+|\mathrm{shift(i_x)}|} - \epsilon & \text{if shift(i}_\text{x}\text{) is defined} \\ 0 & \text{otherwise} \end{array} \right\} \\
S_A & = & \text{Intermediate result when Sequence A is the target sequence} \\
 & = & \displaystyle\sum_{i_A=1}^{N_A} s(i_A) \\
S_B & = & \text{Intermediate result when Sequence B is the target sequence} \\
 & = & \displaystyle\sum_{i_B=1}^{N_B} s(i_B) \\
score & = & \dfrac{S_A + S_B}{|R| + |C|}
\end{array}
$$

where ϵ is a small number with the effect of differentiating minor shift errors from major ones.

The $s(i_x)$ terms are positive for small shifts or shifts of zero. For large shifts, they approach $-\epsilon$. Table 3.1 shows the relation between $|shift(i_x)|$ and $s(i_x)$ as a function of ϵ and for $\epsilon = 0.2$. Residues shifted by no more than one turn of a helix contribute to the score, while residues shifted by more than one turn of a helix decrease the score. In general, if $\epsilon = \frac{1}{n}$, $s(i_x)$ is greater than zero for $|shift(i_x)| < n$ and negative for $|shift(i_x)| > n$.

The shift score is a number between $-\epsilon$ and 1.0. If the two alignments are identical, their shift score is 1.0. If one alignment is a subalignment of the other, then all the $s(i_x)$ terms are 1 for the residue pairs aligned in the subalignment and 0 for the residue pairs excluded from the subalignment. The shift score is related to coverage, as shown:

$$
\begin{array}{rcl}
A & = & \text{the larger alignment} \\
B & = & \text{the subalignment}
\end{array}
$$

$\|\text{shift}(\text{i}_\text{x})\|$	$s(i_x)$	$s(i_x)$ with $\epsilon = 0.2$
0	1	1
1	$\frac{1}{2} - \frac{\epsilon}{2}$	0.4
2	$\frac{1}{3} - \frac{2\times\epsilon}{3}$	0.2
3	$\frac{1}{4} - \frac{3\times\epsilon}{4}$	0.1
4	$\frac{1}{5} - \frac{4\times\epsilon}{5}$	0.04
5	$\frac{1}{6} - \frac{5\times\epsilon}{6}$	0
6	$\frac{1}{7} - \frac{\epsilon}{7}$	-0.0285
7	$\frac{1}{8} - \frac{\epsilon}{8}$	-0.05
8	$\frac{1}{9} - \frac{\epsilon}{9}$	-0.0667
9	$\frac{1}{10} - \frac{\epsilon}{10}$	-0.08
10	$\frac{1}{11} - \frac{\epsilon}{11}$	-0.0909
∞	0 - ϵ	-0.2

Table 3.1: Illustration of the relation between absolute shift and shift score term $s(i_x)$ as a function of ϵ and for $\epsilon = 0.2$.

$$\begin{aligned}
\text{cov} &= \text{fraction of residue pairs aligned in } A \text{ that are also aligned in } B \\
\text{score} &= \text{shift score} \\
&= \frac{2\,|B|}{|A| + |B|} \\
&= \frac{2cov}{1 + cov}
\end{aligned}$$

The shift score does not depend on the assignment of target and template sequence, or reference and candidate alignment. Because the shift score includes coverage information, it does not need to be viewed in the context of coverage or other alignment statistics.

4. Examples

In this chapter, we will conclude by showing some sample alignment shifts, their shift scores, and the shift score for their *optimal subalignment*, the subalignment of the candidate alignment that maximizes the shift score. These figures depict the structure-structure alignment generated by DALI versus a predicted alignment generated with no structural information. The alignment generated by DALI is used as the reference alignment, and is shown. The candidate alignment is not shown, but is reflected in the shift lines. All shift scores shown below are computed with $\epsilon = 0.2$.

In Figure 4.1, the CASP2 target T0031 is aligned to the structure 1try. Much of the alignment is good, as represented by the large number of shifts of zero. At the same time, there are a few positions with shifts as large as 18. When these positions of very large shift are omitted, the shift score improves radically.

Figures 4.2 and 4.3 reflect two different candidate alignments in comparison with the same reference alignment. The alignment reflected in Figure 4.2 is slightly better, and this alignment is given the higher shift score.

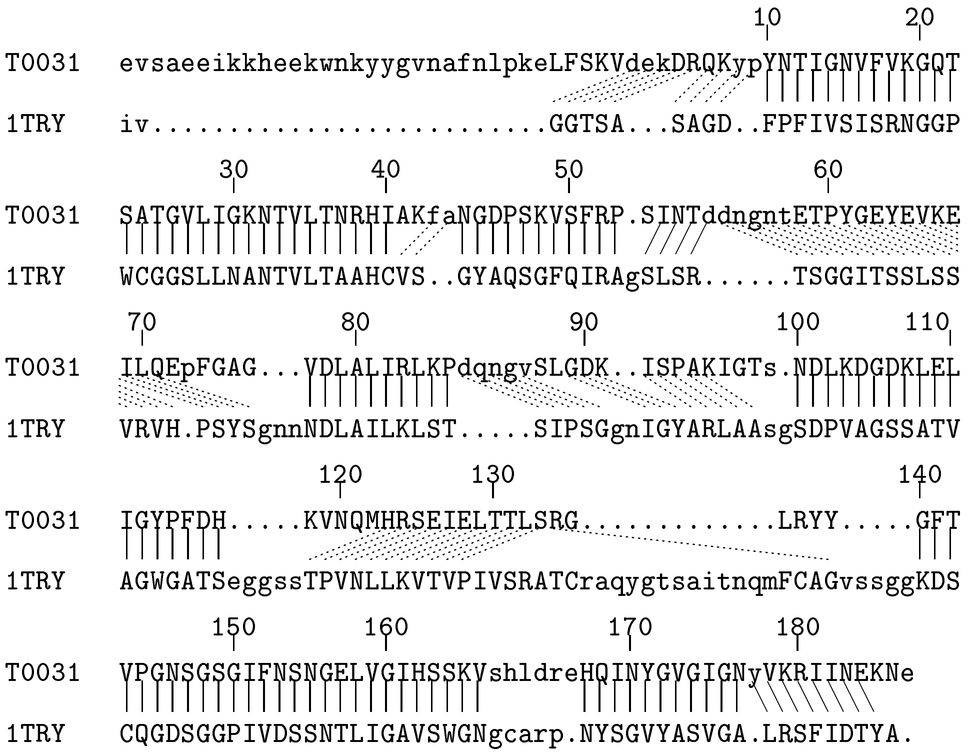

10
20
T0031 evsaeeikkheekwnkyygvnafnlpkeLFSKVdekDRQKypYNTIGNVFVKGQT
1TRY iv.........................GGTSA...SAGD..FPFIVSISRNGGP
30
40
50
60
T0031 SATGVLIGKNTVLTNRHIAKfaNGDPSKVSFRP.SINTddngntETPYGEYEVKE
1TRY WCGGSLLNANTVLTAAHCVS..GYAQSGFQIRAgSLSR......TSGGITSSLSS
70
80
90
100
110
T0031 ILQEpFGAG...VDLALIRLKPdqngvSLGDK..ISPAKIGTs.NDLKDGDKLEL
1TRY VRVH.PSYSgnnNDLAILKLST.....SIPSGgnIGYARLAAsgSDPVAGSSATV
120
130
140
T0031 IGYPFDH.....KVNQMHRSEIELTTLSRG.............LRYY.....GFT
1TRY AGWGATSeggssTPVNLLKVTVPIVSRATCraqygtsaitnqmFCAGvssggKDS
150
160
170
180
T0031 VPGNSGSGIFNSNGELVGIHSSKVshldreHQINYGVGIGNyVKRIINEKNe
1TRY CQGDSGGPIVDSSNTLIGAVSWGNgcarp.NYSGVYASVGA.LRSFIDTYA.

Figure 4.1: Shift score 0.594057. Final shift score 0.719489 achieved by removing from the candidate alignment the columns indicated with dotted lines

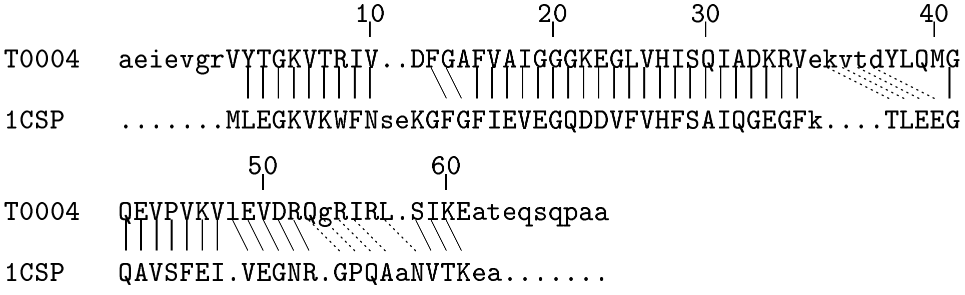

10
20
30
40
T0004 aeievgrVYTGKVTRIV..DFGAFVAIGGGKEGLVHISQIADKRVekvtdYLQMG
1CSP ....MLEGKVKWFNseKGFGFIEVEGQDDVFVHFSAIQGEGFk....TLEEG
50
60
T0004 QEVPVKVlEVDRQgRIRL.SIKEateqsqpaa
1CSP QAVSFEI.VEGNR.GPQAaNVTKea.......

Figure 4.2: Shift score 0.715807. Final shift score 0.761062 achieved by removing from the candidate alignment the columns indicated with dotted lines

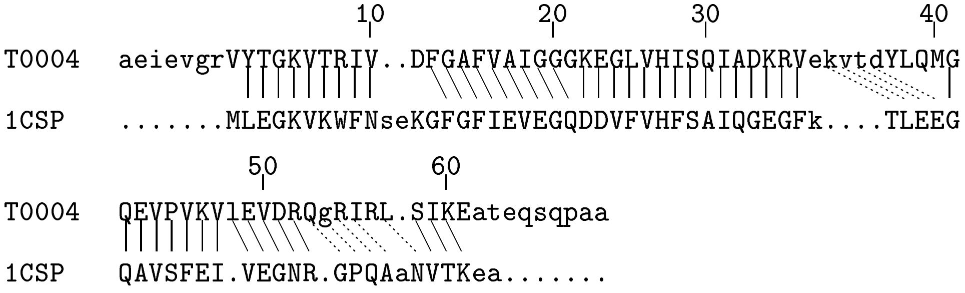

10
20
30
40
T0004 aeievgrVYTGKVTRIV..DFGAFVAIGGGKEGLVHISQIADKRVekvtdYLQMG
1CSP ....MLEGKVKWFNseKGFGFIEVEGQDDVFVHFSAIQGEGFk....TLEEG
50
60
T0004 QEVPVKVlEVDRQgRIRL.SIKEateqsqpaa
1CSP QAVSFEI.VEGNR.GPQAaNVTKea.......

Figure 4.3: Shift score 0.648065. Final shift score 0.686726 achieved by removing from the candidate alignment the columns indicated with dotted lines

References

[Bry96] S. Bryant. Evaluation of threading speicificity and accuracy. *Proteins: Structure, Function, and Genetics*, 26(2):172–85, 1996.

[God96] A. Godzik. The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, 5(7):1325–38, July 1996.

[MBB97] A. Marchler-Bauer and S. Bryant. Measures of threading specificity and accuracy. *Proteins: Structure, Function, and Genetics*, Supplement 1(1):134–139, 1997.