*ISSN 0006-3509, Biophysics, 2008, Vol. 53, No. 4, pp. 253–255. © Pleiades Publishing, Inc., 2008.*
*Original Russian Text © V.O. Polyanovsky, M.A. Roytberg, V.G. Tumanyan, 2008, published in Biofizika, 2008, Vol. 53, No. 4, pp. 533–537.*

# A New Approach to Assessing the Validity of Indels in Algorithmic Pair Alignments

**V. O. Polyanovsky**[a], **M. A. Roytberg**[b], **and V. G. Tumanyan**[a]

[a] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia*
[b] *Institute of Mathematical Problems in Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia*

Received June 18, 2008

**Abstract**—Analysis of the structure of indels in algorithmic versus evolutionary alignments based on a set of inequalities confirms the conclusions from numerical modeling. For the more divergent sequences (PAM > 60), the tested aligning algorithm (SW) tends to increase the mean length of indels and decrease their number.

*Key words*: pairwise alignment, insertion, deletion

**DOI:** 10.1134/S0006350908040015

## INTRODUCTION

The sequences of symbols in biopolymers can be compared using various algorithms of pairwise alignment [1–6]. Applying any procedure, it is important to have an idea of the quality of the aligning algorithm, i.e., of how precisely the algorithmic alignment (obtained by optimization of a certain target function) can reproduce the "evolutionary alignment" of the amino acid (or nucleotide) sequences. The evolutionary alignment implies collation of the positions in the compared polymers that originate from one and the same position of their common ancestor.

Here, the "evolutionarily true" (ET) alignment is based on the data on the disposition of insertions/deletions and substitutions introduced into sequences [7, 8] in an evolutionary model [9, 10].

In a recent work [8] we used numerical modeling to assess the quality of the global version of the Smith–Waterman (SW) algorithm and compare the structure of indels and changes in SW vs. ET alignments. We found that in the algorithmic alignments the overall length of indels was markedly smaller than in the ET ones, because the number of indels decreased considerably while the mean indel length remained the same or slightly increased [8].

Here we present an alternative theoretical approach, whereby the indel characteristics for algorithmic and true alignments can be compared by their mean values without statistical processing of test sets of alignments.

## THE METHOD

**Basic relationships.** Consider the following ratios for indel characteristics in the $i$-th pair of algorithmic (alg) and the evolutionarily true (et) alignments:

$\nu_i = n^i_{\mathrm{alg}}/n^i_{\mathrm{et}}$ for the number of indels,

$\lambda_i = l^i_{\mathrm{alg}}/l^i_{\mathrm{et}}$ for the sum length of indels,

$\alpha_i = \lambda_i/\nu_i$ for the mean length of an indel.

Then the corresponding averages over the set of alignments are

$$\nu_{\mathrm{av}} = \frac{1}{N}\sum_{i=1\ldots N}\nu_i,$$

$$\lambda_{\mathrm{av}} = \frac{1}{N}\sum_{i=1\ldots N}\lambda_i,$$

$$\alpha_{\mathrm{av}} = \frac{1}{N}\sum_{i=1\ldots N}\lambda_i/\nu_i.$$

where $N$ is the number of aligned pairs.

Now compare the sum length of indels with the product of mean indel length by the number of indels:

$$\lambda_{\mathrm{av}} \vee \nu_{\mathrm{av}}\alpha_{\mathrm{av}}$$

or

$$(1/N)\sum_{i=1\ldots N}\lambda_i \vee (1/N^2)\left(\sum_{i=1\ldots N}\nu_i\right)\left(\sum_{i=1\ldots N}\lambda_i/\nu_i\right), \quad (1)$$

whereby the difference is

$$\Delta \equiv (1/N)\sum_{i=1\ldots N}\lambda_i - (1/N^2)\left(\sum_{i=1\ldots N}\nu_i\right)\left(\sum_{i=1\ldots N}\lambda_i/\nu_i\right). \quad (2)$$

It is easy to see that

**Table 1**

| | | |
|---|---|---|
| $\Delta > 0$ | $\sigma_\nu/\nu_{av} > \sigma_\lambda/\lambda_{av}$ | – |
| $\Delta > 0$ | $\sigma_\nu/\nu_{av} < \sigma_\lambda/\lambda_{av}$ | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} > 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} < 0$ |
| $\Delta < 0$ | $\sigma_\nu/\nu_{av} > \sigma_\lambda/\lambda_{av}$ | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} > 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} < 0$<br>$\Delta\nu_{ij} > 0, \Delta\lambda_{ij} < 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} > 0$ |
| $\Delta < 0$ | $\sigma_\nu/\nu_{av} < \sigma_\lambda/\lambda_{av}$ | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} < 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} > 0$ |

**Table 2**

| Length | PAM [9] | Product αν [8] |
|---|---|---|
| 200 | 60 | 0.88 |
| 200 | 100 | 0.84 |
| 200 | 200 | 0.77 |
| 200 | 300 | 0.77 |
| 500 | 60 | 0.92 |
| 500 | 100 | 0.89 |
| 500 | 200 | 0.82 |
| 500 | 300 | 0.82 |

$$\text{(i)}\left(\sum_{i=1\dots N}\nu_i\right)\left(\sum_{i=1\dots N}\alpha_i\right) =$$

$$= \sum_{i=1\dots N}\lambda_i + \sum_{i=1\dots N-1}\;\sum_{j=i+1\dots N}(\nu_i\alpha_j + \nu_j\alpha_i),$$

$$\text{(ii)}\ N\sum_{i=1\dots N}\lambda_i = \sum_{i=1\dots N}\lambda_i + \sum_{i=1\dots N-1}\;\sum_{j=i+1\dots N}(\lambda_i + \lambda_j).$$

and (2) can be written as

$$\Delta = (1/N^2)\sum_{i=1\dots N-1}\;\sum_{j=i+1\dots N}\delta_{ij},$$

where $\delta_{ij} = (\nu_j\lambda_i - \nu_i\lambda_j)(\nu_i - \nu_j)/\nu_i\nu_j$.

**Dependence of the sign of Δ on the combination of signs for differences $\Delta\nu_{ij} = \nu_i - \nu_j$ and $\Delta\lambda_{ij} = \lambda_i - \lambda_j$.** Note that the sign of $\delta_{ij}$ is determined by the combination of inequality signs in (a) $\Delta\nu_{ij}/\nu_j \vee \Delta\lambda_{ij}/\lambda_j$, (b) $\Delta\nu_{ij} \vee 0$, and (c) $\Delta\lambda_{ij} \vee 0$ where $\vee$ can be > or <. Thus there are eight combinations of signs, whereby

$$(\Delta\nu_{ij}/\nu_j >< \Delta\lambda_{ij}/\lambda_j,\ \Delta\nu_{ij} >< 0,\ \Delta\lambda_{ij} >< 0) \Rightarrow \delta_{ij} < 0, \quad (3)$$

$$(\Delta\nu_{ij}/\nu_j >< \Delta\lambda_{ij}/\lambda_j,\ \Delta\nu_{ij} >< 0,\ \Delta\lambda_{ij} >< 0) \Rightarrow \delta_{ij} < 0, \quad (4)$$

$$(\Delta\nu_{ij}/\nu_j >< \Delta\lambda_{ij}/\lambda_j,\ \Delta\nu_{ij} >< 0,\ \Delta\lambda_{ij} >< 0) \Leftrightarrow \delta_{ij} > 0. \quad (5)$$

Since

$$\delta_{ij} >< 0, \quad i, j = 1, \dots, N \Rightarrow \Delta >< 0,$$

$\delta_{ij}$ in (3)–(5) can be replaced with Δ (also replacing ⇔ with ⇒).

**Dependence of the relation between $\sigma_\nu/\nu_{av}$ and $\sigma_\lambda/\lambda_{av}$ on the relation between $(\Delta\nu_{ij})^2$ and $(\Delta\lambda_{ij})^2$.** Since $\nu_{av}, \lambda_{av} > 0$, then $\sigma_\nu/\nu_{av} >< \sigma_\lambda/\lambda_{av} \sim \sigma_\nu\lambda_{av} >< \sigma_\lambda\nu_{av}$.

For rmsd we have

$$\sigma_a = (1/N)\left(\sum_{i=1\dots N-1}\;\sum_{j=i+1\dots N}\Delta a_{ij}^2\right)^{1/2}.$$

Then

$$\sigma_a c_{av} =$$

$$= (1/N)^2\left[\left(\sum_{i=1\dots N-1}\;\sum_{j=i+1\dots N}\Delta a_{ij}^2\right)\left(\sum_{i=1\dots N}b_i\right)^2\right]^{1/2},$$

where $a = \nu, \lambda$, $b = \lambda, \nu$.

Since

$$\left(\sum_{i=1\dots N}b_i\right)^2 = \sum_{i=1\dots N}b_i\Bigg)^2 + 2\sum_{i=1\dots N-1}\;\sum_{j=i+1\dots N}b_ib_j$$

we can state that

$$\begin{aligned}&\forall i, j, k, l = 1, \dots, N\colon (\Delta\nu_{ij})^2\lambda_k\lambda_l >\\ &<(\Delta\lambda_{ij})^2\nu_k\nu_l \Rightarrow \sigma_\nu\lambda_{av} >< \sigma_\lambda\nu_{av}\end{aligned} \quad (6)$$

or:

if for all $i, j, k, l = 1, \dots, N$ it is true that

$$(\Delta\nu_{ij})^2\lambda_k\lambda_l >< (\Delta\lambda_{ij})^2\nu_k\nu_l,$$

then $\sigma_\nu\lambda_{av} >< \sigma_\lambda\nu_{av}$.

Assuming that

$$\begin{aligned}&((\forall i, j, k, l = 1, \dots, N\colon (\Delta\nu_{ij})^2\lambda_k\lambda_l >< (\Delta\lambda_{ij})^2\nu_k\nu_l\\ &\Leftrightarrow \forall i, j, k, l = 1, \dots, N\colon |\Delta\nu_{ij}|\lambda_j >< |\Delta\lambda_{ij}|\nu_j,\end{aligned}$$

from (3)–(6) we get

$$\begin{aligned}&(\Delta\nu_{ij}/\nu_j >< \Delta\lambda_{ij}/\lambda_j,\ \Delta\nu_{ij} <> 0,\ \Delta\lambda_{ij} <> 0)\\ &\Rightarrow \{\Delta > 0,\ \sigma_\nu/\nu_{av} < \sigma_\lambda/\lambda_{av}),\end{aligned} \quad (7)$$

$$\begin{aligned}&(\Delta\nu_{ij}/\nu_j >< \Delta\lambda_{ij}/\lambda_j,\ \Delta\nu_{ij} >< 0,\ \Delta\lambda_{ij} >< 0)\\ &\Rightarrow \{\Delta < 0,\ \sigma_\nu/\nu_{av} > \sigma_\lambda/\lambda_{av}),\end{aligned} \quad (8)$$

$$\begin{aligned}&(\Delta\nu_{ij}/\nu_j >< \Delta\lambda_{ij}/\lambda_j,\ \Delta\nu_{ij} >< 0,\ \Delta\lambda_{ij} <> 0)\\ &\Rightarrow \{(\Delta < 0,\ \sigma_\nu/\nu_{av} \vee \sigma_\lambda/\lambda_{av})\end{aligned} \quad (9)$$

which can be presented in tabular form (Table 1).

**Table 3**

| Length | PAM | $\sigma_\nu/\nu_{av}$ | $\sigma_\nu/\lambda_{av}$ | $\Delta$ | Combination |
|---|---|---|---|---|---|
| 200 | 60 | 0.1929 | 0.2184 | –0.0031 | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} < 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} > 0$ |
| 200 | 100 | 0.2061 | 0.2384 | –0.0057 | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} < 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} > 0$ |
| 200 | 200 | 0.2628 | 0.3033 | –0.0124 | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} < 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} > 0$ |
| 200 | 300 | 0.3311 | 0.3619 | –0.0255 | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} < 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} > 0$ |
| 500 | 60 | 0.1211 | 0.1449 | 0.0046 | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} > 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} < 0$ |
| 500 | 100 | 0.1355 | 0.162 | –0.0006 | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} < 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} > 0$ |
| 500 | 200 | 0.1895 | 0.2036 | –0.0284 | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} < 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} > 0$ |
| 500 | 300 | 0.2547 | 0.238 | –0.0118 | $\Delta\nu_{ij} > 0, \Delta\lambda_{ij} < 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} > 0$<br>$\Delta\nu_{ij} > 0, \Delta\lambda_{ij} > 0$<br>$\Delta\nu_{ij} < 0, \Delta\lambda_{ij} < 0$ |

## RESULTS AND DISCUSSION

Using the averaged $\nu$, $\lambda$, $\alpha$ values from our recent work [8], we can obtain the required (mean length by number) product (Table 2) and analyze the relative characteristics of indels for various evolutionary distances as regards the trends in sign combinations (Table 3).

Non-coincidence of the $\Delta\nu_{ij}$ and $\Delta\lambda_{ij}$ signs (rightmost column) means that a higher (lower) ratio of the indel numbers goes together with a lower (higher) ratio of the mean indel lengths (i.e., indel lengths are not the same in the algorithmic and the ET alignments). Judging by the increase in the $\Delta$ absolute value, this tendency is pronounced in all cases but two: at $L = 500$ and PAM = 60 the signs agree and the mean indel lengths may be equal; at $L = 500$ and PAM = 300 the assessment is inconclusive. In any case, this analysis shows that for the more divergent sequences (less than 55% identity) the increase of the sum length of indels in the algorithmic alignment is caused by overestimation of the mean indel length and underestimation of the indel number as compared with the true alignment. That is, the SW algorithm of pairwise alignment tends to alter the structure of indels.

## REFERENCES

1. V. G. Tumanyan, L. E. Sotnikova, and A. V. Kholopov, Dokl. RAN **166**, 1465 (1966).
2. V. V. Poroikov, N. G. Esipova, and V. G. Tumanyan, Mol. Biol. **18**, 541 (1984).
3. T. F. Smith and M. S. Waterman. J. Mol. Biol. **147**, 195 (1981).
4. S. B. Needleman and C. D. Wunsch. J. Mol. Biol. **48**, 443 (1970).
5. S. F. Altschul, W. Gish, W. Miller, et al., J. Mol. Biol. **215**, 403 (1990).
6. D. J. Lipman and W. R. Pearson, Science **227**, 1435 (1985).
7. V. O. Polyanovsky, E. Ya. Demchuk, and V. G. Tumanyan, Mol. Biol. **28**, 1341 (1994).
8. V. O. Polyanovsky, M. A. Roytberg, and V. G. Tumanyan, J. Comput. Biol. **15**, 379 (2008).
9. M. Dayhoff, R. Schwartz, and B. Orcutt, in *A model of evolutionary change in proteins*, Ed. M. Dayhoff, *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, 1978), pp. 345–352.
10. S. A. Benner, M. A. Cohen, and G. H. Gonnet, J. Mol. Biol. **229**, 1065 (1993).