Sequence Weights

Stephen F. Altschul

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

The Problem

Given an alignment $\mathcal{A}$ of several sequences (we will defer until later how to construct such an alignment), how should we define scores for aligning $\mathcal{A}$ to a single new sequence?

More simply, leaving aside the question of gap scores, how should one score the alignment of a multiple alignment column $\mathcal{C}$ to a single letter?

This problem raises three distinct, and deep questions:

How does one deal with correlation among the aligned sequences?

How many independent observations does an alignment column represent?

How does account for small sample size and for prior knowledge?

We will defer the third question until later.

[It is] as if someone were to buy several copies of the morning paper to assure himself that what it said was true.

Ludwig Wittgenstein (1953) *Philosophical Investigations*, part I, §265.

Starting Intuitions

Counting all sequences equally can lead to a loss of information when a sequence is copied multiple times, because it can dilute independent information from other sequences. Identical or nearly identical copies of the same sequence provide little new information.

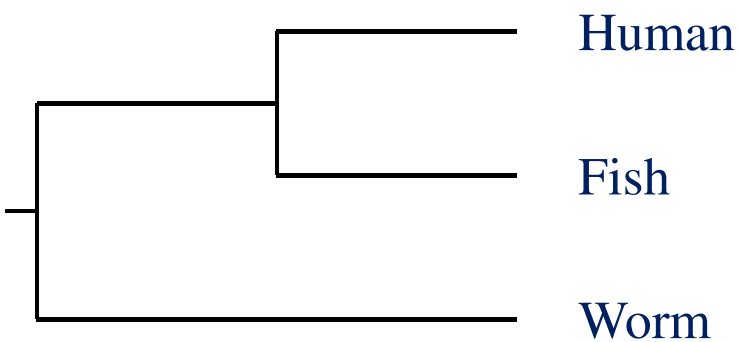
It may be possible to mitigate this problem by giving each sequence a *weight*, with nearly identical sequences downweighted, and unusual sequence upweighted.

Questions:

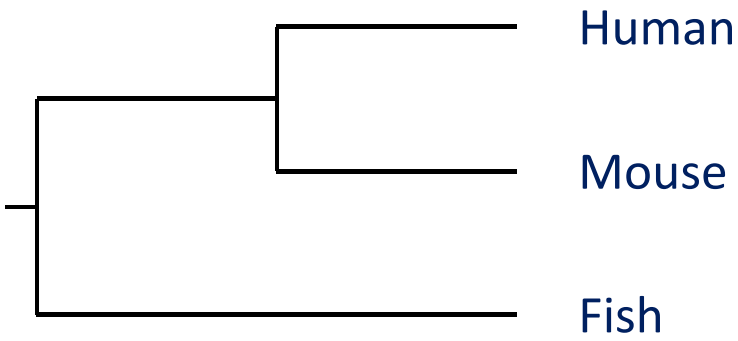
How can one formalize this problem?

Can one recast the problem of finding appropriate sequence weights as an optimization problem?

Weights Depend on a Set of Sequences



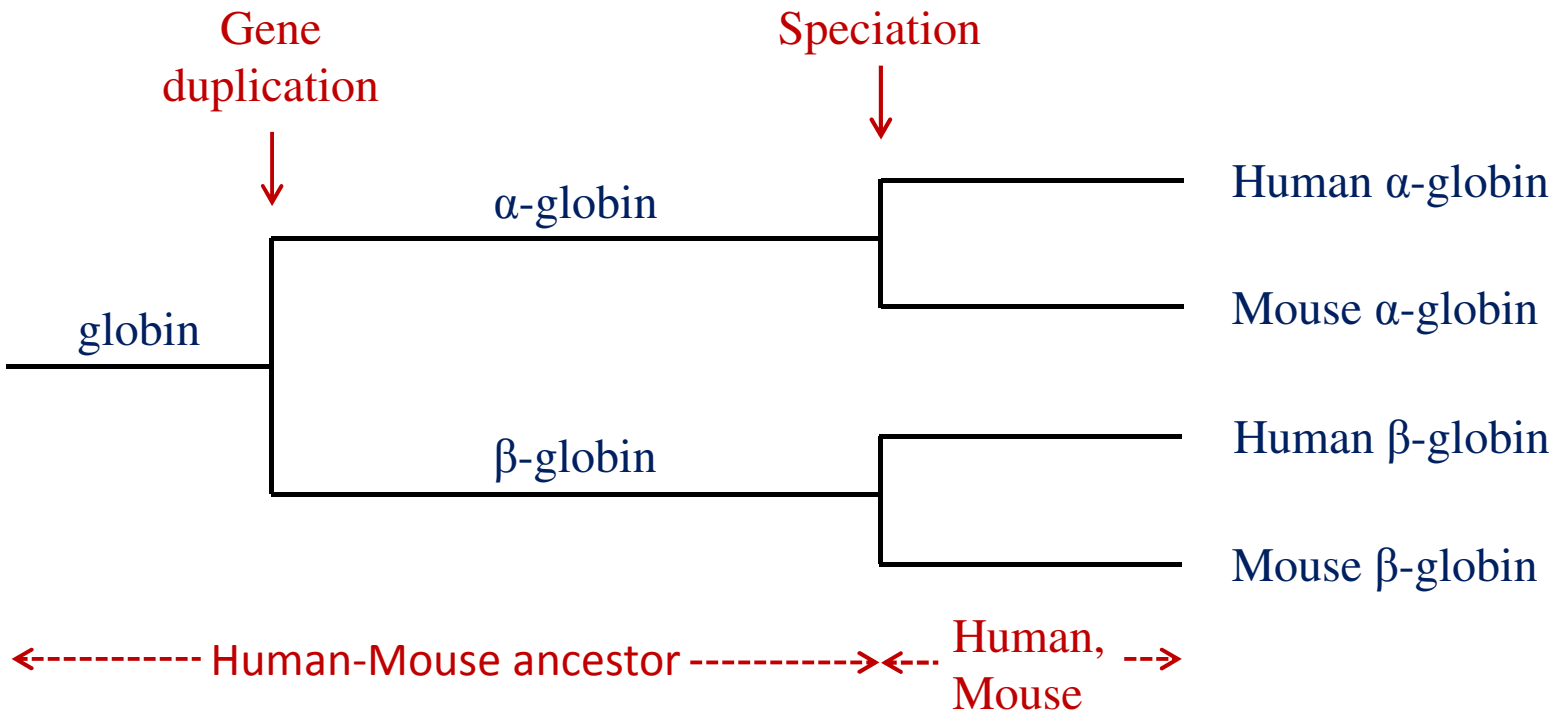
Given orthologous genes or proteins from these three organisms, the fish sequence should be downweighted.



Given orthologous genes or proteins from these three organisms, the fish sequence should be upweighted.

In other words, a weight is never *intrinsic* to a sequence. It is associated with a sequence only in the context of a set of other sequences.

Digression: Orthology and Paralogy

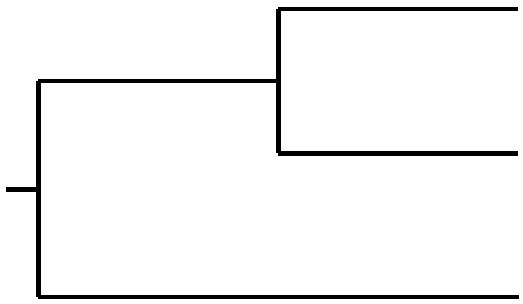


Homology: Two genes or proteins are *homologous* if they share a common ancestor.

Orthology: Two genes or proteins are *orthologous* if they diverged by speciation.

Paralogy: Two genes or proteins are *paralogous* if they diverged by gene duplication.

Sequence Trees and Phylogenetic Trees

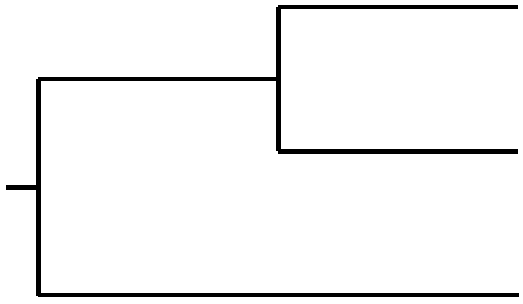


Human α-globin

Mouse α-globin

Fish α-globin

For orthologous genes or proteins, a tree reflecting sequence relationships should be congruent with the *phylogentic tree* of species relationships.



Human α-globin

Fish α-globin

Mouse β-globin

For paralogous genes or proteins, a tree reflecting sequence relationships may be incongruent with the phylogentic tree of species relationships.

Over the course of evolution, it is possible that in a particular protein family different paralogs are lost in different species. In that case there may be no set of orthologs for that family from which a valid phylogenetic tree may be reconstructed.

Literature

There have been many approaches to the weighting problem, including:

Felsenstein, J. (1973) *Am. J. Hum. Genet.* **25**:471-492.

Felsenstein, J. (1985) *Am. Nat.* **125:**1-15.

Altschul, S.F., *et al.* (1989) *J. Mol. Biol.* **207**:647-653.

Sibbald, P.R. & Argos, P. (1990) *J. Mol. Biol.* ***2*16**:813-818.

Sander C. & Schneider, R. (1991) *Proteins* **9**:56-68.

Vingron, M. & Sibbald, P.R. (1993) *Proc. Natl. Acad. Sci. USA* ***9*0**:8777-8781.

Gerstein, M., *et al.* (1994) *J. Mol. Biol*. **236**:1067-1078.

Henikoff, S. & Henikoff, J.G. (1994) *J. Mol. Biol*. **243**:574-578.

Thompson, J.D., *et al.* (1994) *Comput. Appl. Biosci.* **10**:19-29.

Eddy, S.R., *et al.* (1995) *J. Comput. Biol.* **2**:9-23.

Gotoh, O. (1995) *Comput Appl. Biosci*. **11**:543-551.

Krogh, A. & Mitchison, G. (1995) In *Proc. Third Int. Conf. on Intelligent Systems for Mol. Biol.* (C. Rawlings *et al*., eds.) pp. 215-221, AAAI Press, Menlo Park, CA.

Bailey, T.L. & Gribskov, M. (1996) In *Proc. Fourth Int. Conf. on Intelligent Systems for Mol. Biol.* (D.J. States *et al*., eds.) pp. 15-24, AAAI Press, Menlo Park, CA.

Sunyaev, S.R., *et al.* (1999) *Protein Eng*. **12**:387-394.

Method A: Purging

A simple approach to dealing with sequence correlation is simply removing or ignoring sequences that are more than *X*% identical to some sequence already included.

Advantages:

- Very fast and simple.
- Duplicating a sequence does not alter results.

Disadvantages:

- No definition of what is being optimized.
- Dependent on order in which sequences are considered.
- Some information is clearly lost.

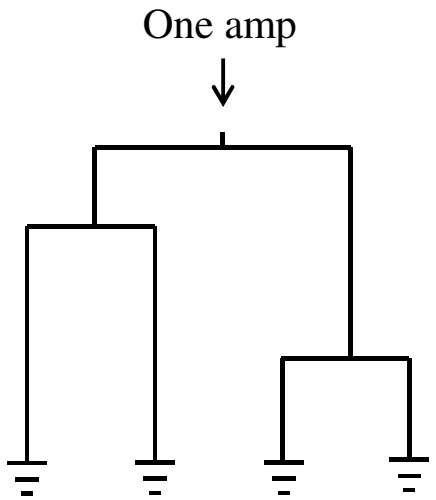
Note: To evaluate correlation among the sequences involved, weighting methods in general rely upon alignments having an appreciable number of columns.

Method B: Tree-Based Weights

Reformulation: Let ***T*** be a continuous one-dimensional quantitative trait that undergoes Brownian motion over the course of evolution. Assume it has value t at the root of a tree, and the values $\vec{t}$ at the tree's leaves.

Question: Given $\vec{t}$, what is the maximum-likelihood estimator for t?

Solution: Let $l_{i,i}$ be the distance from the root to leaf i, and let $l_{i,j}$ be the distance from the root to the last common ancestor of leaves i and j. Then the variance of the random variable t_i is proportional to $l_{i,i}$, and the covariance of t_i and t_j is proportional to $l_{i,j}$. Let $\mathbf{M}$ be the variance-covariance matrix, $\vec{1}$ be a column vector of 1s, and $\vec{w} = (\mathbf{M}^{-1}\vec{1})/(\vec{1}'\mathbf{M}^{-1}\vec{1})$. Then $\hat{t} = \vec{w} \bullet \vec{t}$ is the estimator we seek.



Equivalent to: Make the vertical edges of the tree of resistant wire, and ground the leaves. Apply a voltage so that one amp flows into the root. The current that flows out each leaf is the weight for that leaf.

Felsenstein, J. (1973) "Maximum-likelihood estimation of evolutionary trees from continuous characters." *Am. J. Hum. Genet.* **25**:471-492.

Altschul, S.F., *et al.* (1989) "Weights for data related by a tree." *J. Mol. Biol.* **207**:647-653.

Tree-Based Weights continued

Advantages:

- Well-formulated as an optimization problem.
- Independent of sequence order.
- Uses all information.
- Tree may be *rooted* anywhere, allowing outgroups to contribute.

Possible disadvantages:

- Leaves farther from the root are downweighted.
- Assumes an evolutionary tree relating the sequences.

Major disadvantage:

- Requires the construction of an evolutionary tree, a hard and time-consuming problem.

Method C: Henikoff Weights

Central idea:

Averaged over multiple-alignment columns, a sequence this is similar to others will tend to have many letters in common with those sequences.

Method:

i) For each column, divide a total weight of 1 evenly among the letter types that occur at that position, and then divide the weight assigned to each letter type evenly among the sequences that have that letter.

ii) For each sequence, sum its weights from all positions, and normalize.

Example:

Sequences	Calculation			Weight
G C G T T A G C	¼ + ⅓ + ⅓ + ¼ + ¼ + ⅓ + ¼ + ½	=	2½	0.31250
G A G T T G G A	¼ + ⅓ + ⅓ + ¼ + ¼ + ⅓ + ¼ + ¼	=	2¼	0.28125
C G G A C T A A	½ + ⅓ + ⅓ + ½ + ½ + ⅓ + ½ + ¼	=	3¼	0.40625

Henikoff, S. & Henikoff, J.G. (1994) "Position-based sequence weights." *J. Mol. Biol.* **243**:574-578.

Henikoff Weights continued

Advantages:

Very fast and simple.

Independent of sequence order.

Uses all information.

Disadvantages:

Ad hoc: no objective function to optimize.

Exact duplication of a sequence does *not* halve its weight. *Why?*

Digression: The Effective Number of Independent Sequences in a Multiple Alignment

Why is this number relevant?

The problem: What, for example, should be the score for aligning a valine to a column of five leucines?

```
...GEALGRLLVVYPWTQ...
...GEALGRLLIVYPWTQ...
...GETLGRLLVVYPWTQ...
...GKALGRLLIVYPWTQ...
...GEALGRLLVVYPWTQ...

..........V..........
```

Here, the sequences in the multiple alignment are virtually identical. There is little reason to score the alignment much differently than that of valine to a single leucine (BLOSUM-62: +1).

```
...GEALGRLLVVYPWTQ...
...KECFTKFLSAHHDIA...
...VVFYTSILEKAPAAK...
...VDILVKFLTGTPAAQ...
...AEGLERTLHSFPTTK...

..........V..........
```

Here, the sequences are very different, providing good evidence that a leucine is highly favored at this position. Thus, the score for aligning a valine should probably be negative.

Estimating the Effective Number of Independent Sequences

Assume the background probabilities of the amino acids are $\vec{p}$.

Given a column of n random, independently chosen amino acids, the expected number of *distinct* amino acids it contains is:

$$f(n) = 20 - \sum_{i=1}^{20}(1 - p_i)^n.$$

Note that f may be extended to real n, and is monotonically increasing.

Thus, for a multiple alignment whose columns have, on average, A distinct amino acids, one may estimate the number of "independent sequences" it represents as $f^{-1}(A)$.

Altschul, S.F., *et al*. (2009) "PSI-BLAST pseudocounts and the minimum description length principle." *Nucleic Acids Res.* **37**:815-824.

This question is addressed as well in:

Altschul, S.F., *et al.* (1997) *Nucleic Acids Res*. **25**:3389-3402.

Sunyaev, S.R., *et al.* (1999) *Protein Eng.* **12**:387-394.

Brown, D.P., *et al.* (2007) *PloS Comput. Biol.* **3**:e160

Method D: PSIC Weights

Central idea:

Weights are used only to estimate the numbers of independent observations for the various letters in a column. Let's estimate these numbers directly.

Method:

i) For a given column c, and a given letter a, first confine attention to the set S of sequences that contain letter a in column c.

ii) Now, ignoring column c, estimate the number of independent sequences in S, using, for example, the mean number of distinct letters in the columns of S.

Example:

2	**1**	**1**	**3**		**2**	**3**	**3**	**2**	**3**	**3**	**2**	**3**	**1**	**2**
L	S	C	F	P	I	S	G	H	L	S	L	Y	W	Y
L	**T**	**C**	**L**	**V**	**V**	**P**	**G**	**D**	**I**	**D**	**V**	**E**	**W**	**T**
L	K	C	Q	Q	N	F	N	N	D	T	M	Y	W	Y
F	**T**	**C**	**F**	**V**	**V**	**L**	**K**	**D**	**A**	**H**	**L**	**T**	**W**	**E**
L	**T**	**C**	**R**	**V**	**S**	**A**	**D**	**G**	**L**	**E**	**V**	**S**	**W**	**E**
L	N	C	T	F	S	S	A	S	Q	Y	F	W	W	Y
L	R	C	K	Y	S	S	A	T	P	Y	L	F	W	Y

$\Rightarrow$ **Mean = 2.21**

$\Downarrow$

2.31 independent observations of valine in column 5.

Calculated by previously described method

Sunyaev, S.R., *et al.* (1999) "PSIC: profile extraction from sequence alignments with position-specific counts of independent observations." *Protein Eng.* **12**:387-394.

PSIC Weights continued

Advantages:

Relatively fast.

Independent of sequence order.

Uses all information.

Duplicating a sequence does not alter results.

At least one independent observation for all observed letters.

Disadvantages:

To some extent *ad hoc*: no objective function to optimize.

Note: Unlike other methods, PSIC does not assign weights to individual sequences. Furthermore, the total number of independent observations it implies varies from one column to another. In some applications, these facts may constitute a disadvantage.