Scoring: Similarity vs. Distance

- Scoring so far:
  - Type 1: similarity (e.g., PAM)
  - Type 2: distances $\hat{=} \text{metric (edit distance)}$

- Often: similarity measurable (PAM, BLOSUM) but distance needed
  
  eg. for computing guide tree for progressive multiple sequence alignment

- However: similarities can be translated into distances!

One of the first approaches by Feng&Doolittle (1987)

- **Given**: 2 sequences $a, b \in \Sigma^*$ and a similarity function $S$
- **Wanted**: (evolutionary) distance estimation $D$ of $a$ and $b$
Feng-Doolittle (1987): Similarity → Distance

- **Definitions:**
  - \( S(a, b) \)  
    similarity of \( a, b \) (= score of opt. alignment)
  - \( S_{a,b}^{\text{max}} = \frac{S(a, a) + S(b, b)}{2} \)  
    maximal similarity possible
  - \( S_{\text{rand}} \)  
    expected score for aligning two random sequences \( a', b' \) with same length and compositions ⇒ \( a, b \) shuffled

- **Then:** define \( S_{a,b}^{\text{eff}} = \frac{S(a, b) - S_{\text{rand}}}{S_{a,b}^{\text{max}} - S_{\text{rand}}} \)

- \( S_{a,b}^{\text{eff}} \) normalized percentage similarity
  - values between 0 and 1
  - converges exponentially slow to 0 for increasing evolutionary distance

- **Hence:** for an approx. linear distance define \( D(a, b) = - \log \left( S_{a,b}^{\text{eff}} \right) \)
Feng-Doolittle (1996): Simplifying $S_{\text{rand}}$ computation

- **Problem 1:** $S_{\text{rand}}$ computationally very demanding
  - Have to:
    - generate random $a', b'$
    - align to get similarity score $S(a', b')$
    - derive statistics and expected score
    - needed for every sequence pair independently! (no general distance score)

- **Problem 2:** $D = -\log(S_{\text{eff}})$ not deterministic! (due to $S_{\text{rand}}$)

- **Solution:** approximate using single (optimal) alignment $A$ of $a, b \in \Sigma^*$
  
  $$S_{\text{rand}} = \frac{1}{|A|} \cdot \left( \sum_{(x, y) \in \Sigma \times \Sigma} S(x, y) \cdot N_a(x) \cdot N_b(y) \right) + \text{gaps}(A) \cdot S(-, \ast)$$

  $$\approx \text{frequency-weighted similarity score}$$

  $N_a(x) =$ occurrences of $x$ in $a$; $\text{gaps}(A) =$ number of gaps in $A$