Twilight Zone of Protein Sequence Alignments

1. Introduction

1.1 Definition of Terms

- *Pairwise sequence identity* is the percentage of residues identical between two aligned sequences.

- Two ways of defining *pairwise sequence similarity*:

  - (i) similarity score between two sequences normalized by the maximal possible score observed in a given metric, or

  - (ii) percentage of residues similar between two aligned sequences.

1.2 Some results on pairwise sequence identity

- Most protein pairs with more than 30% pairwise sequence identity were found to be structurally similar. (Sander and Schneider, 1991)

- Structure alignments have uncovered homologous protein pairs with less than 10% pairwise sequence identity. (Valencia *et al.*, 1991; Holmes *et al.*, 1993; Holm and Sander, 1996; Brenner *et al.*, 1996; Hubbard *et al.*, 1997)

1.3 The twilight zone

• Sequence alignment methods unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity is high ($>$ 40% for long alignments). This region is refered to as the *safe zone*.

• However, conventional sequence alignment methods often fail to correctly align proteins with 20-30% pairwise sequence identity. Hence, Doolittle (1986) coined the term *twilight zone* for sequence alignments.

• Below 20% pairwise sequence identity, it is even harder to predict protein pairs of similar structure. The author refers to this region as the *midnight zone*.

• Pairwise sequence identity is not sufficient to define the twilight zone.

• In 1990, Sander and Schneider defined a length-dependent threshold for significant sequence identity based on the relatively small number of structure pairs available at that time.

• The threshold curve (HSSP-curve) was defined as follows:

$$t(L) = \begin{cases} 290.15 \cdot L^{-0.562} & \text{for } 10 \leq L \leq 80 \\ 25 & \text{for } L > 80 \end{cases}$$

where L is the number of residues aligned between two proteins (ignoring gaps).

1.4 Goals of the Paper

- (1) Characterize the protein pairs in the twilight zone.

- (2) Determine if the HSSP-curve defined by Sander and Schneider (1991) is still valid for the current protein databases.

- (3) Define a length-dependent threshold for significant sequence similarity and determine if using sequence similarity rather than identity will improve accuracy.

- (4) Finally, determine if accuracy be improved for pair alignments by expert rules.

2. Methods

2.1 Data set used

- Protein databases are biased towards particular protein families. To reduce this bias, analyses are usually restricted to representative data sets (Hobohm *et al.*, 1992).

- This investigation used the maximal set of 792 sequence-unique (no pair in the set falls above the HSSP-curve) proteins of known structure available in early 1997.

- Each of the 792 sequence-unique proteins was aligned against the subset of PDB contained in the early 1997 release of the FSSP database of protein structure alignments. This subset amounted to about 5646 protein chains.

2.2 Generation of sequence alignments

- Protein pairs were aligned by two different program types:

  - (i) Full dynamic programming as implemented in the Smith-Waterman based method MaxHom (Schneider, 1994)

  - (ii) quick database searches as implemented by the two versions of the BLAST series: BLASTP (Altschul et al., 1990; Altschul and Gish, 1996), and PSI-BLAST (Altschul et al., 1997).

- Alignments shorter than 10 residues were not considered, as identical polypeptides of up to 10 residues are known to occur in different structure states.

- Technical limitations (CPU time) required the restriction of the dynamic-programming analysis to the best 2000 hits for each of the 792 unique proteins.

2.3 Standard of truth for structural similarity

- Since different structure alignment methods yield different scores, similarity between two proteins is not uniquely defined.

- Here, the author chose the FSSP database as the 'standard of truth': any pair for which FSSP listed a significant score of structural similarity was considered to be structurally similar.

2.4 Concept of true and false hits

- The following cases were distinguished:

  - (i) *true positives*, alignments between proteins of similar structure that fall above a given threshold (defined by the sequence alignment method)

  - (ii) *false positives*, alignments between proteins of dissimilar structure that fall above a given threshold of the sequence alignment

  - (iii) *true negatives*, alignments between proteins of dissimilar structure that that fall below a given threshold

  - (iv) *false negatives*, alignments between proteins of similar structure that fall below a given threshold

2.5 New curve for length-dependent significance of pairwise sequence identity

- The author revised the original HSSP-curve by defining the following curve for the separation of true and false positives:

$$t^I(L) = 480 \cdot L^{-0.32 \cdot (1+e^{-L/100})}$$

where L gave the number of residues aligned between two proteins, and t^I is the cut-off percentage of identical residues over the L aligned residues.

2.6 New curve for length-dependent significance of pairwise sequence similarity

- The results indicate that the functional dependence between similarity and length appeared comparable to the one between identity and length. This prompted a similar definition for the separation of true and false positives based on similarity:

$$t^S(L) = 420 \cdot L^{-0.335 \cdot (1+e^{-L/2000})}$$

2.7 Sequence-space-hopping

• Sequence-space-hopping refers to combining two protein families that share a common sequence. In this study, it was used to improve detection accuracy by discarding false positives.

2.7 More-similar-than-identical rule

• A simple rule of thumb was explored to distinguish between true and false positives: accept hits only if the level of sequence similarity was higher than the level of sequence identity.