Wilhelm-Schickard-Institut für Informatik (WSI)
Zentrum für BioInformatik Tübingen (ZBIT)
Lehrstuhl Rechnerarchitektur (Prof. Dr. A. Zell)

Eberhard Karls
Universität
Tübingen

Similarity-preserving Metrics for Amino-acid Sequences

String distances

- Levenshtein (edit) distance
- Feature (n-gram) distance
- Negative scaled similarity [Agrafiotis]
- Mole fraction metric [Apostol & Szpankowski]
- Setubal & Meidanis

Problems: either not biologically plausible or mathematically inconsistent.

Levenshtein: minimum number of edit operations needed to transform one string into another. No biological connection.

Feature: number of n-grams in which the strings differ. Biologically unfounded and mathematically inconsistent.

Agrafiotis: based on scoring matrices, but mathematically inconsistent.

Apostol & Szpankowski: takes only mole fractions of amino-acids into account, discards their order.

Setubal & Meidanis: mathematically and biologically sound, but requires all symbols to have the same self-similarity.

Similarity-based metric

Analogy with vectors:

$d(a,b) := \langle a|a\rangle + \langle b|b\rangle - 2\langle a|b\rangle$

- Symmetric, positive-semidefinite, zero for a=b
- For certain scoring schemes (BLOSUM etc.) satisfies the triangle inequality

Distance

- Symmetric, positive-semidefinite measure
- Satisfies the triangle inequaltiy
- In vector spaces computed over scalar product
- Foundation for many algorithms

Examples: values like mean, median, and variance can be defined over a distance measure.

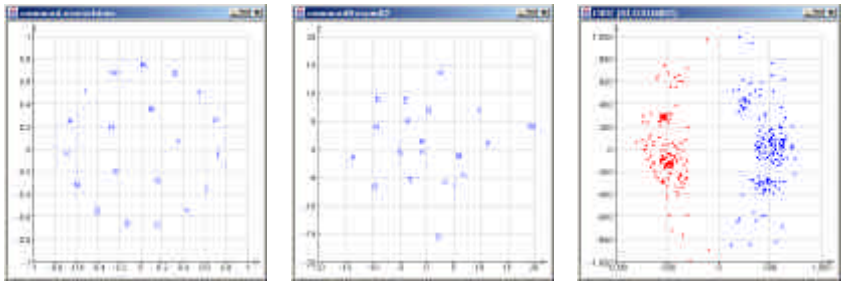
Similarity score

- Biologically founded: PAM, BLOSUM, DNA... schemes
- Mathematical properties undefined
- Not applicable in many statistical and pattern recognition algortihms

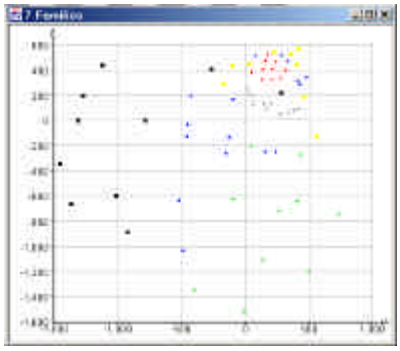
Question: given similarities $\langle a|b\rangle$, $\langle b|c\rangle$ such that $\langle a|b\rangle < \langle b|c\rangle$, what can be said about $\langle a|c\rangle$?

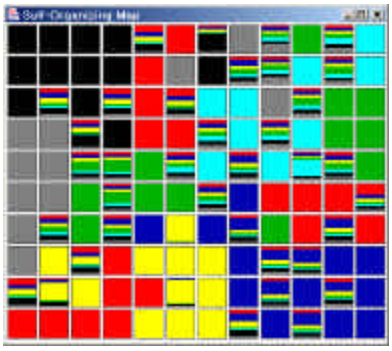
Applications of distance measures for strings

- Sammon Mapping: projecting the data onto a plane with preserving the distances
- Self-Organizing Maps: regression through mapped data



Above: Sammon mappings of the identity and BLOSUM62 matrices and of the hemoglobin α and β chains. **Below:** Sammon mapping and the Self-Organizing Map of samples from seven protein families.





Contact: Igor Fischer, fischer@informatik.uni-tuebingen.de