The limits of protein sequence comparison
William R Pearson and Michael L Sierk
Science 2005

Presentation by: Brook Chernet
Focus of the study

- Homology vs. Analogy
- Homology and statistical significance
- Sequence similarity statistics
- Progress in sequence similarity searching
- Evaluating search algorithms
Introduction

- Emergence of biological sequence comparison programs
- Inference of homology
- Methods of homolog identification
  - Sequence - sequence alignment
  - Sequence – profile alignment
  - Structural alignment
Profile Method
Important Keys

- SCOP – Structural Classification Of Proteins
- CATH - is a hierarchical classification of protein domain structures.
- PDB – protein database bank

<table>
<thead>
<tr>
<th>SCOP</th>
<th>CATH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>Class</td>
</tr>
<tr>
<td>Family</td>
<td>Architecture</td>
</tr>
<tr>
<td>Super Family</td>
<td>Topology</td>
</tr>
<tr>
<td>Fold</td>
<td>Homology</td>
</tr>
</tbody>
</table>

Derived from secondary structure content
Orientation of secondary structures
Topological connections and numbers of secondary structures
Determined by sequence and structural alignment
Homology vs. Analogy

Bovine trypsin

Streptomyces griseus trypsin

S. griseus protease

Viral serine protease

Subtilisin

Homologs, analogs(?) and convergent evolution. Three-dimensional structures of five serine proteases: (a) bovine trypsin (PDB code 5PTP), (b) Streptomyces griseus trypsin (PDB code 1SGT), (c) S. griseus protease A (PDB code 2SGA), (d) viral serine protease (PDB code 1BEF) and (e) subtilisin (PDB code 1SBT). The CATH structure classification places 5PTP, 1SGT and 2SGA in the same homology category, whereas 1BEF has the same topology, but is classified as non-homologous to 5PTP. SCOP places 1BEF in the same superfamily as 5PTP. Subtilisin (1SBT) has a very different structure to the trypsin-like serine proteases and is clearly non-homologous. However, the active sites of subtilisin and trypsin are examples of convergent evolution.
- Trypsin-like serine proteases belong to the mainly-B class of CATH

- Structural similarity (B-barrel) among 1a-c

- Subtilisins exhibit d/t overall 3D structure with the same catalytic function at its active site – Convergent evolution

- Inference of homology based on degree of similarity and how unlikely that two structures arisen independently

- Measure of statistical significance
- DALI, VAST are structure-based comparison applications
- PSI-BLAST and COMPASS performs sequence-profile comparison
- SSEARCH does pairwise sequence alignment
Sequence similarity statistics

- The need to base the inference of homology on statistics
- Use of sequence, structure and function to determine homology
  - Accuracy
  - 30-40% sequence alignment threshold
  - Structure function relationship
Progress in sequence similarity searching

- Karlin-Altschul algorithm of BLAST
- Smith-Waterman algorithm of SSEARCH
- Searching a sequence against sets of aligned sequences
  - Hidden Markov Models (HMMs)
  - Position specific scoring matrices (PSSMs)
  - More sensitive
  - PFAM – Profile database
Evaluating search algorithms

Figure 2: Homologous Protein Coverage – SSEARCH (25%), PSI-BLAST (40%), COMPASS (60%), DALI (98%), VAST (70%).
Conclusions

■ Structural comparison > Profile-sequence comparison > sequence-sequence comparison.

■ Profile methods are important in identifying distant relationships.

■ Excessive similarity (i.e. similar structure, function and sequence) leads to the inference of homology.
Future Improvements

- Revise search algorithms to account for conserved regions
- Build super-super-super computers to run these database searches.
THANK YOU!