The limits of protein sequence comparison

William R Pearson and Michael L Sierk

Science 2005

Presentation by: Brook Chernet

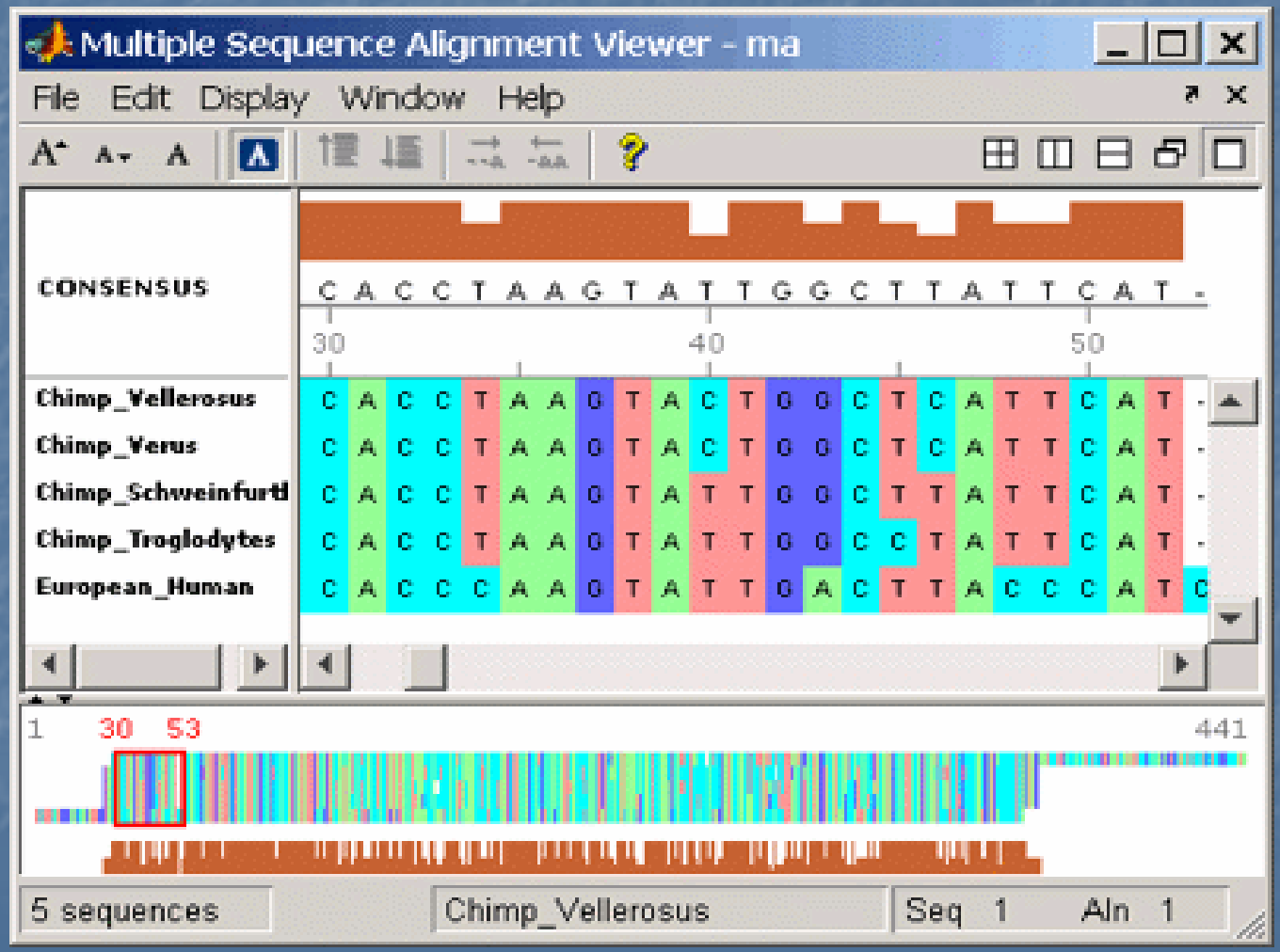
Focus of the study

- 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

SCOP	CATH
Class	Class
Family	Architecture
Super Family	Topology
Fold	Homology

Derived from secondary structure content

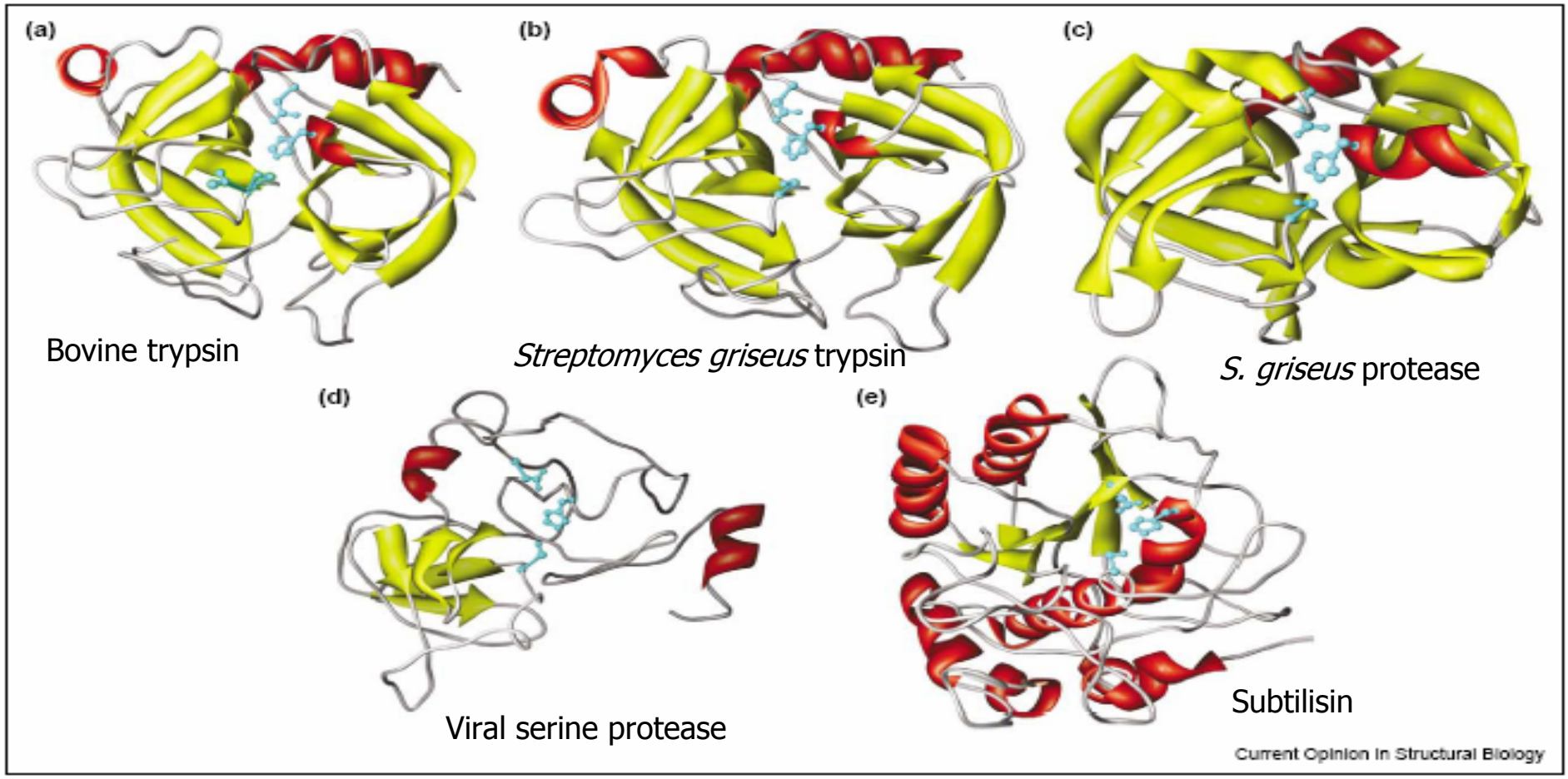
Orientation of secondary structures

Topological connections and numbers of secondary structures

Determined by sequence and structural alignment

Homology vs. Analogy

Figure 1 Trypsin-like serine protease



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
- Subtilisin exhibits 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

Table 1

Similarity for homologs and non-homologs[a].

	5PTP versus:	1SGT	2SGA	1BEF	1SBT
DALI	Z	32.7	13.7	8.8	<2
	E(2775)	10^{-14}	10^{-4}	0.02	>100
	N_{align} (%id)	209 (34)	147 (19)	131 (10)	NA
	RMSD (Å)	1.4	2.8	2.9	NA
VAST	E(2775)	10^{-21}	0.017[b]	1.94	NA
	N_{align} (%id)	208 (34)	130 (22)	122 (14)	NA
	RMSD (Å)	1.5	2.3	2.8	NA
COMPASS	E(10 000)	10^{-114}	10^{-13}	0.056	13
PSI-BLAST	E(2775)	10^{-48}	2.5	>10	>10
	N_{align}	231	40	NA	NA
SSEARCH	E(10 000)	10^{-19}	2.6	>10	>10
	N_{align} (%id)	223 (36)	181 (25)	68 (33)	159 (25)

[a] N_{align} is the number of aligned residues, %id is the percent sequence identity and RMSD is the root mean square distance. E(N) is the expectation value, the number of times a score is expected by chance in a search of a database of size N. 5PTP, 1SGT and 2SGA are trypsin-like serine protease homologs; 1BEF is a viral protease that CATH says is not homologous to the trypsin-like proteases, but SCOP says is homologous to them. 1SBT is subtilisin, which has the same catalytic triad as the trypsin-like serine proteases, but a completely different global domain structure.

[b] Based on alignment with 5SGA, which is 100% sequence identical to 2SGA. NA: not available — no alignment was calculated.

- 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

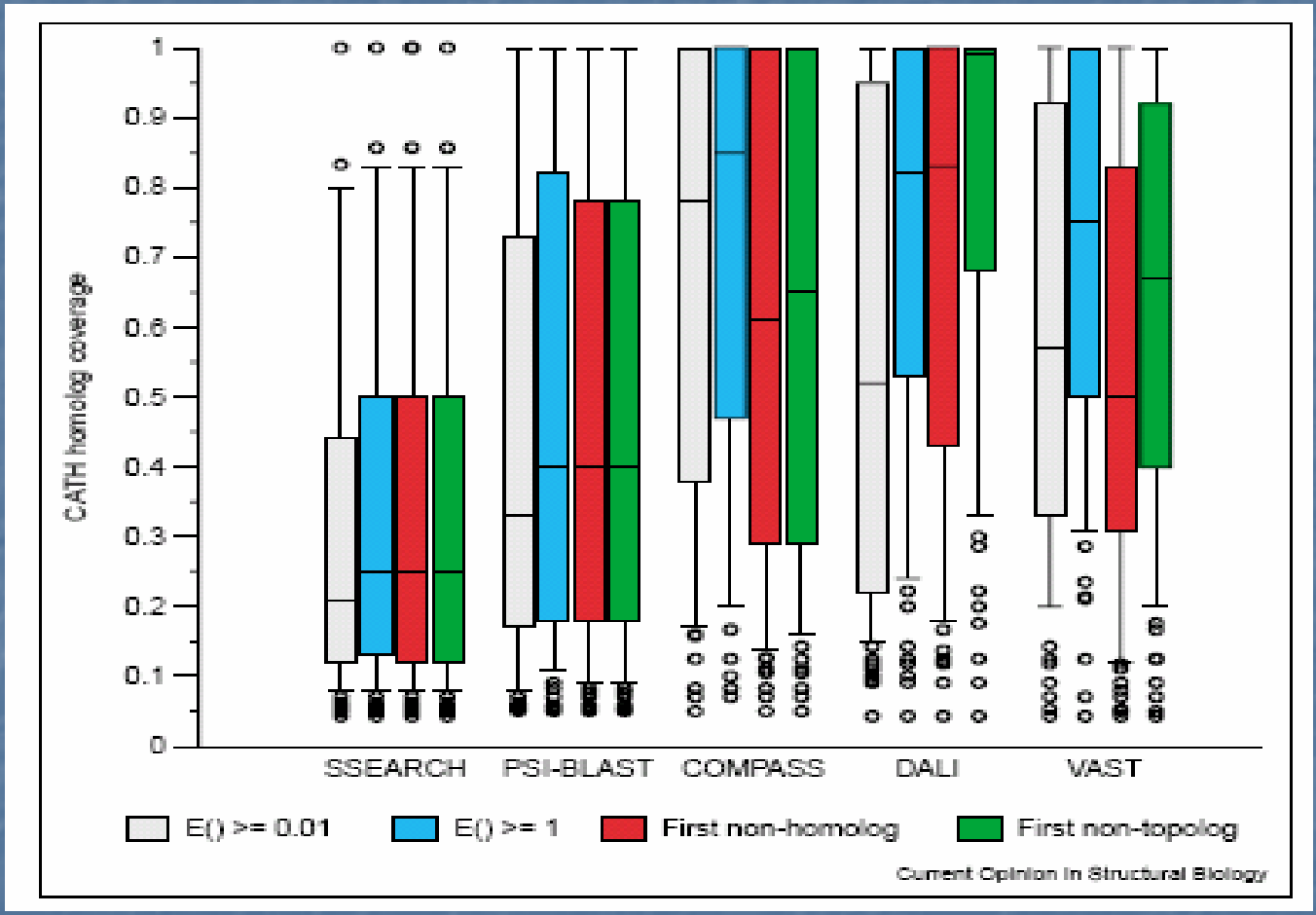


Figure 2: Homologous Protein Coverage – SSEARCH (25%), PSI-BLAST (40%) COMPASS (60%), DALI (98%), VAST (70%).

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!