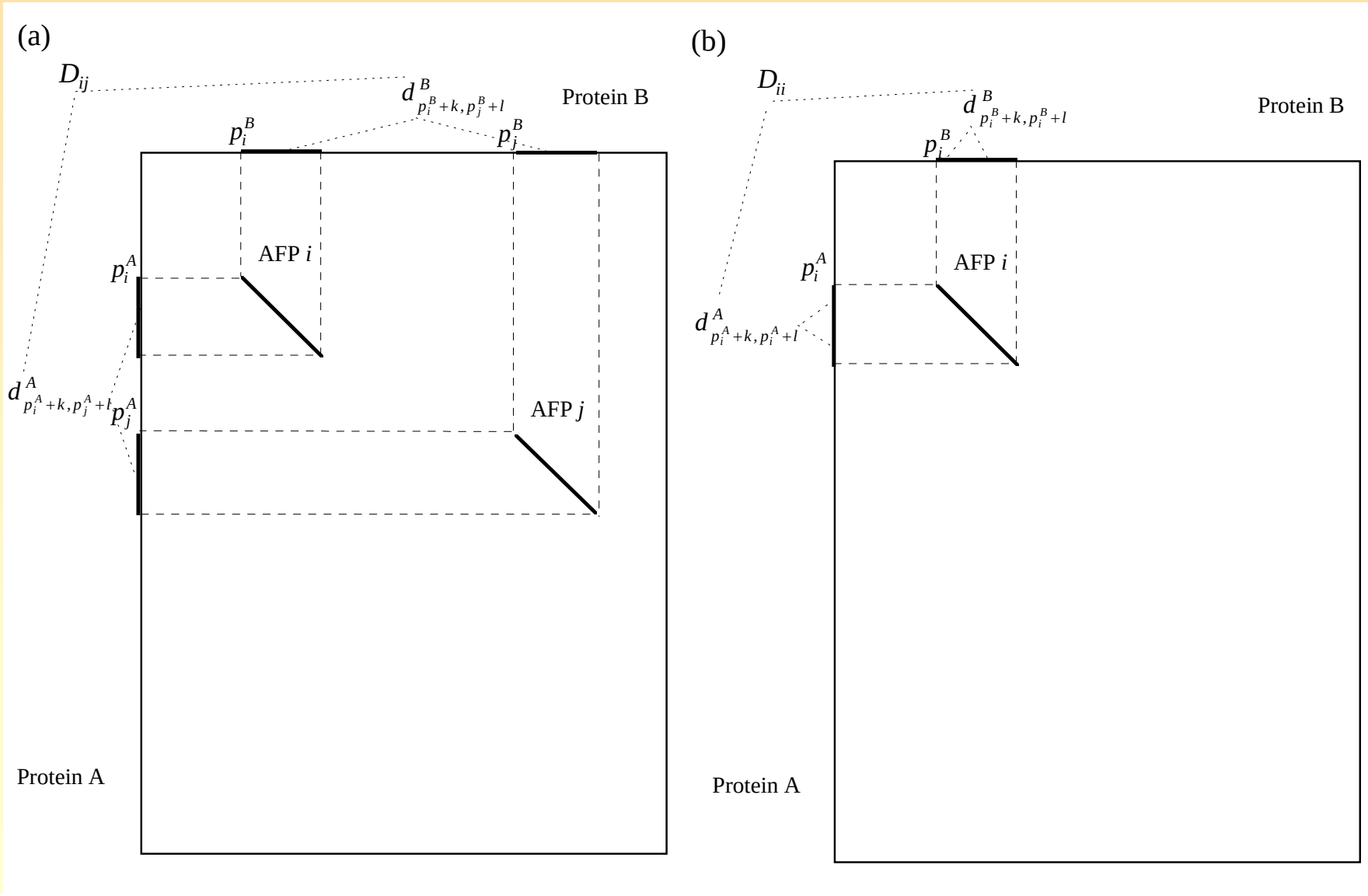
Understand one method (CE) in more detail

I.N. Shindyalov and P.E. Bourne *Protein Engineering* 1998, **11(9)** 739-747. Protein Structure Alignment by Incremental Combinatorial Extension of the Optimum Path. [PDF File]

Basic Approach

- Compare octameric fragments – an aligned fragment pair (AFP)
- Stitch together AFPs
- Find the optimal path through the AFPs
- Optimize the alignment through dynamic programming
- Measure the statistical significance of the alignment

Calculation of distance: (a) D_{ij} for alignment represented by two AFPs i and j from the path; (b) D_{ii} for single AFP i from the path.



Definition of the Alignment Path

$$p_{i+1}^{A} = p_i^{A} + m \text{ and } p_{i+1}^{B} = p_i^{B} + m \quad (1)$$

or

$$p_{i+1}^{A} > p_i^{A} + m \text{ and } p_{i+1}^{B} = p_i^{B} + m \quad (2)$$

or

$$p_{i+1}^{A} = p_i^{A} + m \text{ and } p_{i+1}^{B} > p_i^{B} + m \quad (3)$$

$$p_{i+1}^{A} \leq p_i^{A} + m + G \quad (4)$$

and

$$p_{i+1}^{B} \leq p_i^{B} + m + G \quad (5)$$

Evaluation based upon the following three distance similarity measures

1. Distance calculated from independent set of inter-residue distances where each distance is used only once

2. Full set of inter-residue distances

3. Rmsd from least squares superposition

Evaluation based upon the following three distance similarity measures

1. Distance calculated from independent set of inter-residue distances where each distance is used only once - used for combinations of 2 AFPs

2. Full set of inter-residue distances - used for a single AFP

3. Rmsd from least squares superposition - used to select few best fragments

How to Extend the Path?

1. Consider all possible AFPs that extend the path

2. Consider only the best AFP

3. Use some intermediate strategy

How to Extend the Path?

1. Consider all possible AFPs that extend the path
   Computationally expensive

2. Consider only the best AFP
   Works well with the right heuristics

3. Use some intermediate strategy

What Heuristics?

$$D_{nn} < D_0 \qquad (8)$$

$$\frac{1}{n-1}\sum_{i=0}^{n-1} D_{in} < D_1 \qquad (9)$$

$$\frac{1}{n^2}\sum_{i=0}^{n}\sum_{j=0}^{n} D_{ij} < D_1 \qquad (10)$$

Candidate AFPs are based upon 8
The best AFP is based upon 9
The decision to extend or terminate the path is based upon 10

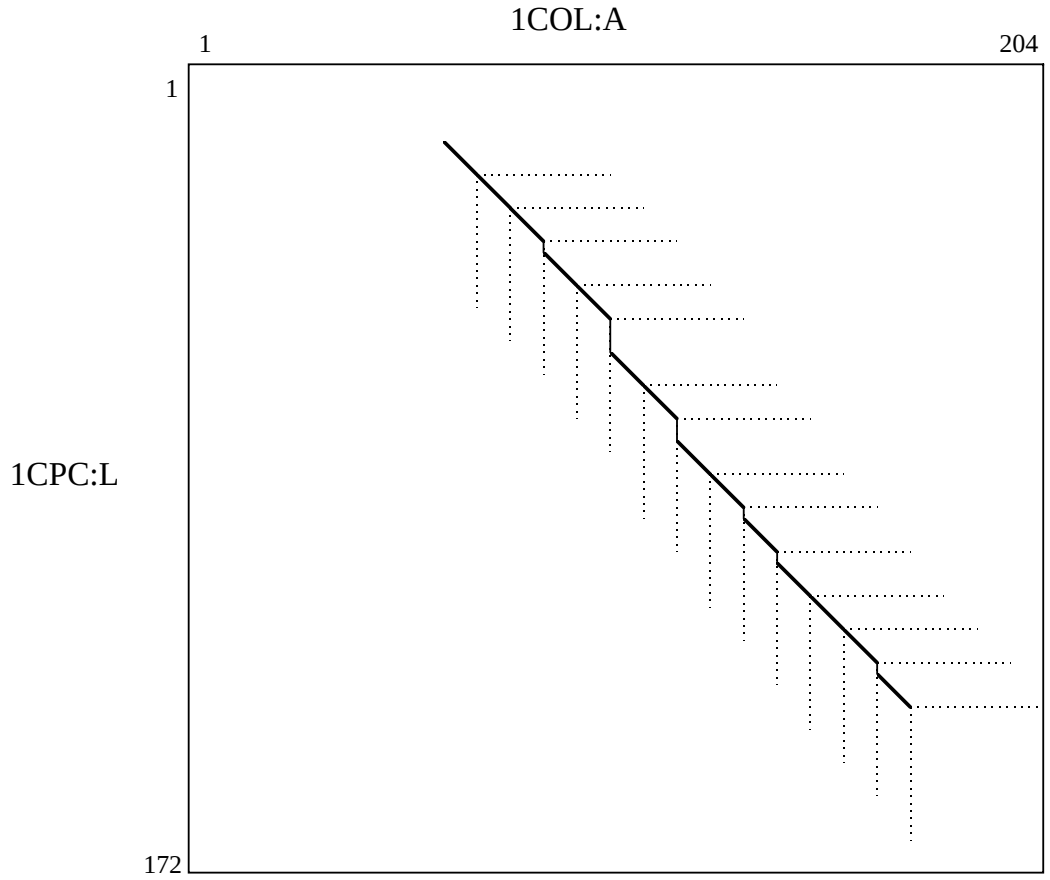
Optimization of the Final Path

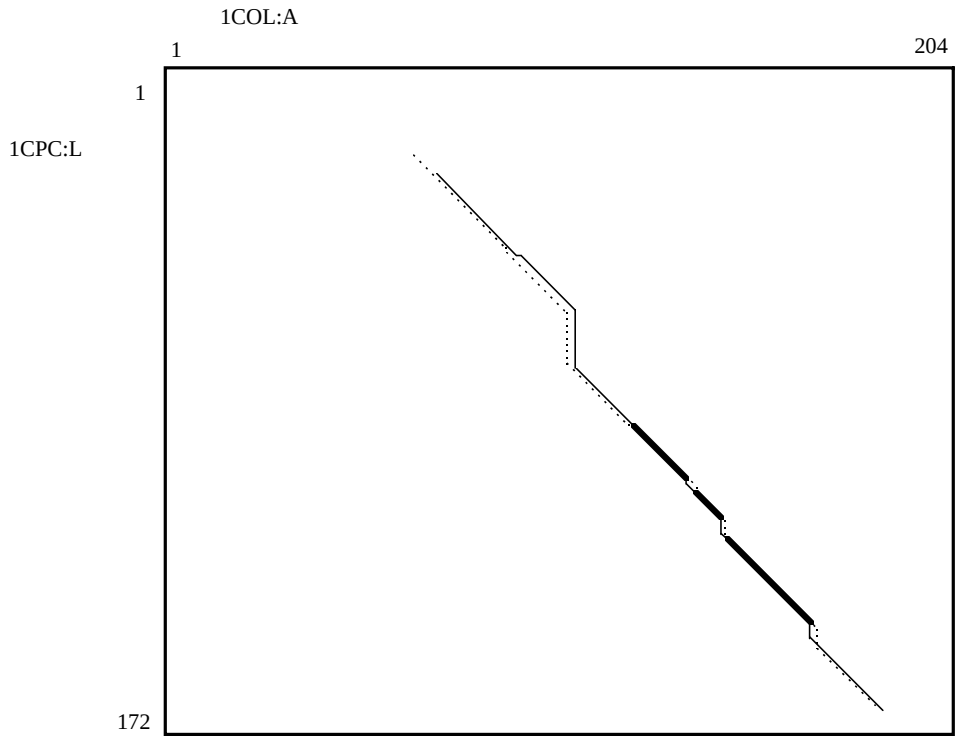
The 20 best alignments with a Z score above 3.5 are assessed based on rmsd and the best kept. This produces approx. one error in 1000 structures

Each gap in this alignment is assessed for relocation up to $m/2$

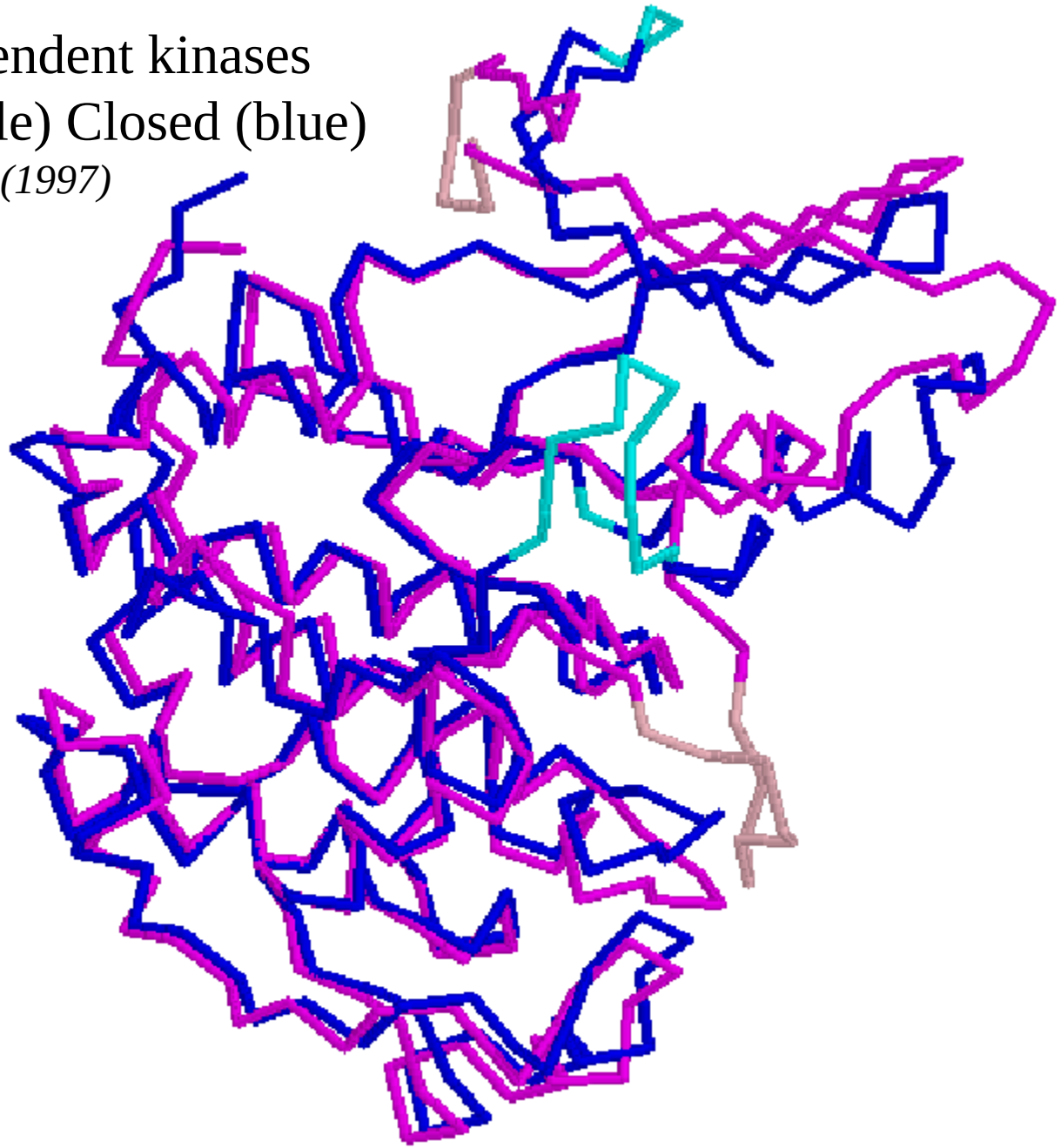
Iterative optimization using dynamic programming is performed using residues for the superimposed structures

Test Case: Phycocyanin versus Colicin A

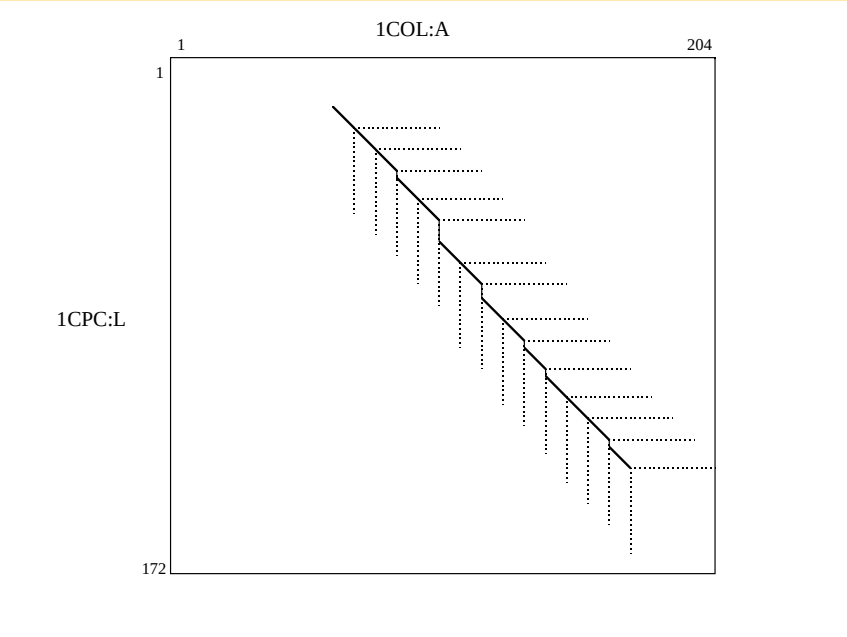



1COL:A
1
204
1
1CPC:L
172

Cyclin-dependent kinases
Open (purple) Closed (blue)
Pavelitch et al. (1997)



Limitations



- Will not find non-topological alignments (outside the bounds of the dotted lines)
- What are the correct “units” to be comparing?
- CE works on chains – domains are the correct units, but definition of the domains is not straightforward

Computation of All x All

- Took 11,748 chain in the PDB (1/98)
- Computed for 1868 representatives
- 24,000 Cray T3E processor hours
- Loaded pairwise alignments into database

2004

- 27,000 proteins ~ 45,000 chains
- $45{,}000^2/2$ * 30 seconds = 963 yrs
- Options:
  - Use a redundant set of chains
  - Use parallel architectures

D. Pekurovsky, I.N. Shindyalov, P.E. Bourne 2004 High Throughput Biological Data Processing on Massively Parallel Computers. A Case Study of Pairwise Structure Comparison and Alignment Using the Combinatorial Extension (CE) Algorithm. *Bioinformatics, 20(12) 1940-1947* [PDF].

One Criterion for Redundancy

- Remove highly homologous chains;
- The RMSD between two chains is less than 2Å;
- The length difference between two chains is less than 10%;
- The number of gap positions in alignment between two chains is less than 20% of aligned residue positions;
- At least 2/3 of the residue positions in the represented chain are aligned with the representing chain.

Review examples where structure comparison has revealed new biological insights

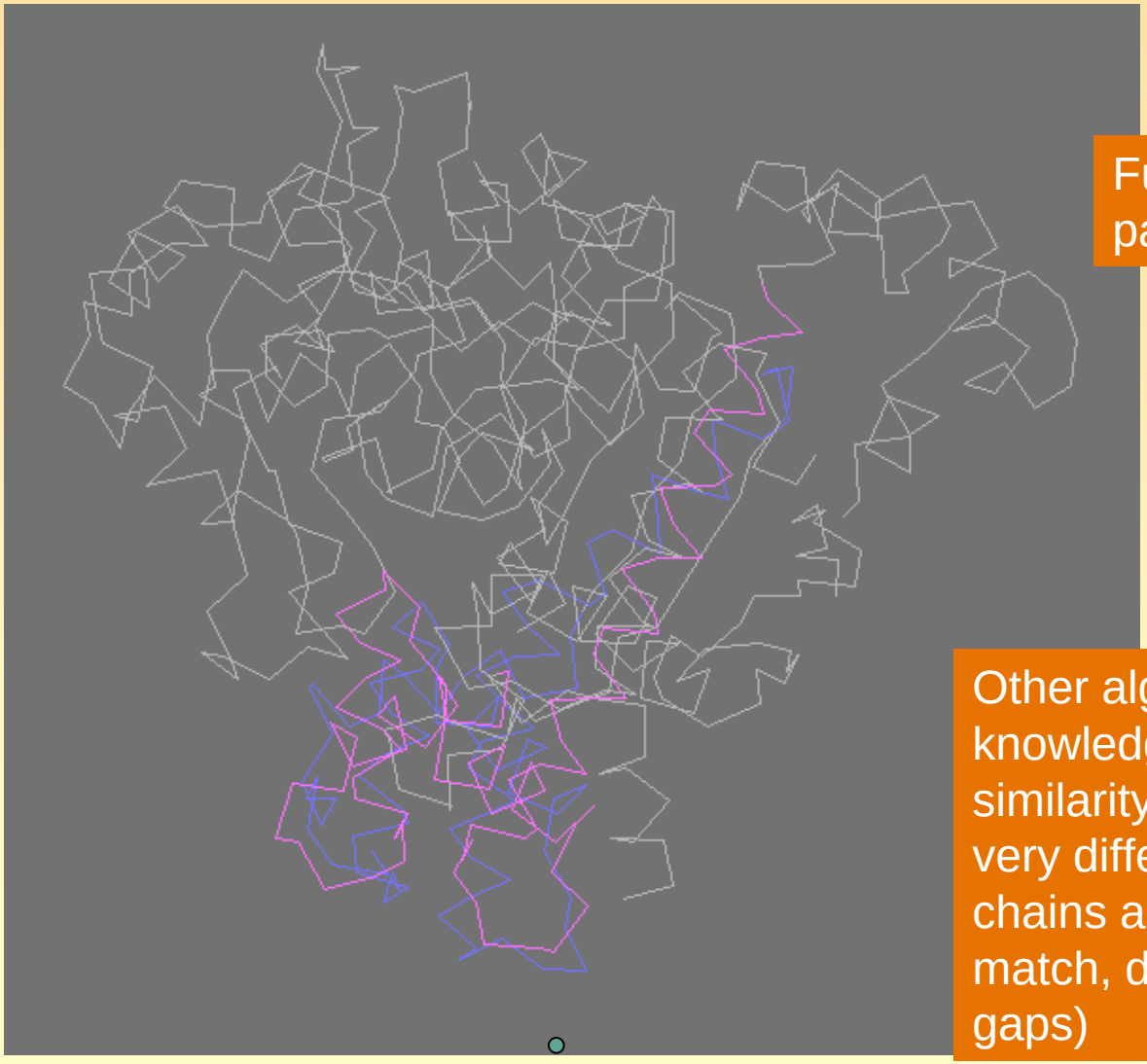
Example 1

Structural similarity between Acetylcholinesterase and Calmodulin found using CE (Tsigelny et al, *Prot Sci*, 2000, **9**:180)

- CE revealed putative Ca++ binding domain in acetylcholine esterase
- Sequence similarity to neuroligins predicts Ca++ binding too – confirmed experimentally
- Members of the a/b hydrolase family bind Ca++ which may be important for heterologous cell associations

Example 1- cont.

2ACE vs. 1TN4: RMSD = 4.6A Z-score = 4.6 LALI = 86 LGAP = 8 Seq. Identity = 3.5%



Full view – nonaligned parts are in grey

Other algorithms to author's best knowledge cannot find this similarity - possibly because of very different size (537 vs. 159) of chains and high RMSD of the match, despite low number of gaps)

Example 2

NCBI

PubMed

National Library of Medicine NLM

Entrez PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for Go Clear

Limits Preview/Index History Clipboard Details

Display Abstract Show: 20 Sort Send to Text

About Entrez

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health

1: Eur J Biochem. 2004 Aug;271(15):3155-70.

Related Articles, Links

Full text blackwell-synergy.com

Structural and functional analysis of ataxin-2 and ataxin-3.

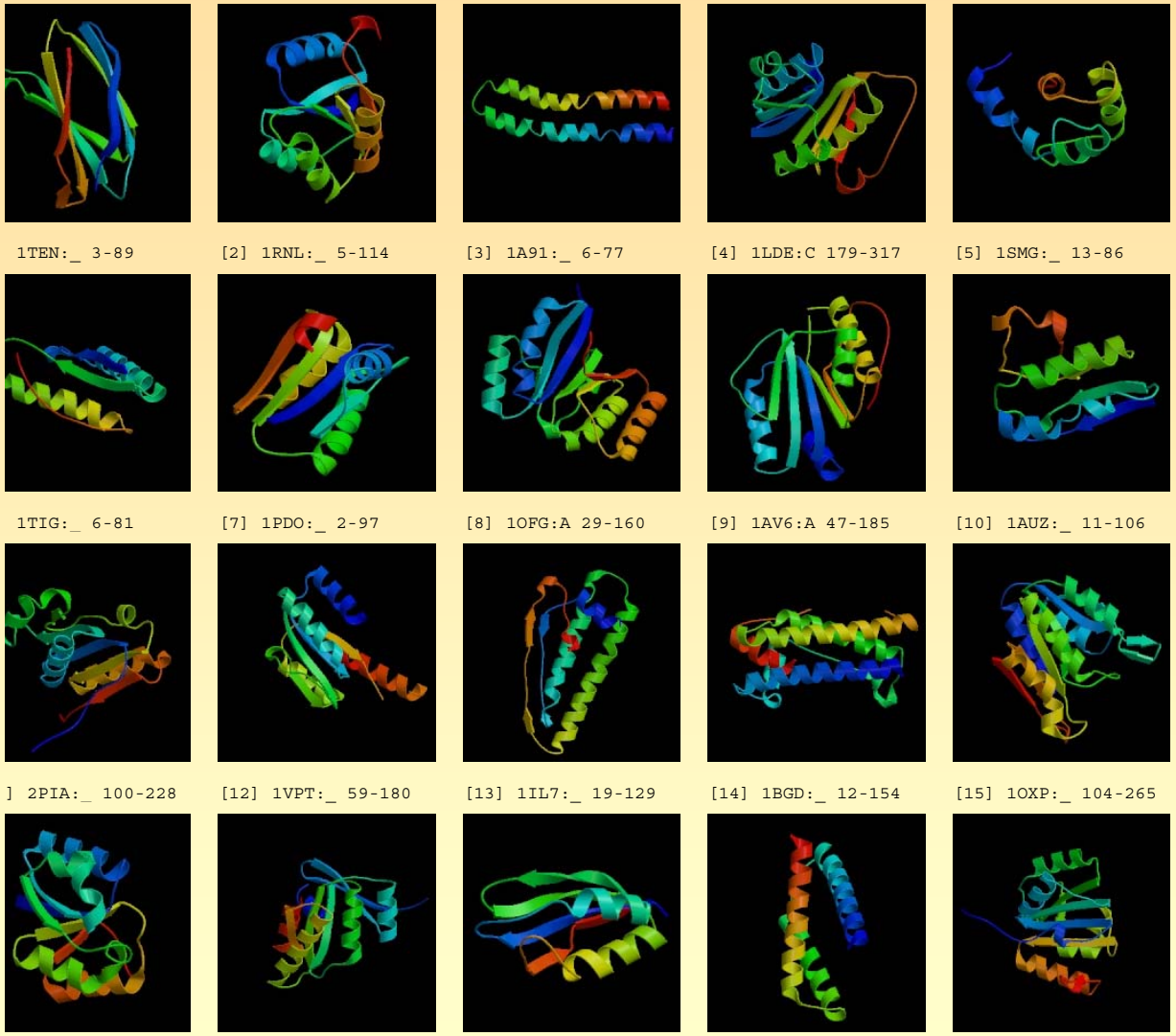
Albrecht M, Golatta M, Wullner U, Lengauer T.

Max-Planck-Institute for Informatics, Saarbrucken, Germany. mario.albrecht@mpi-sb.mpg.de

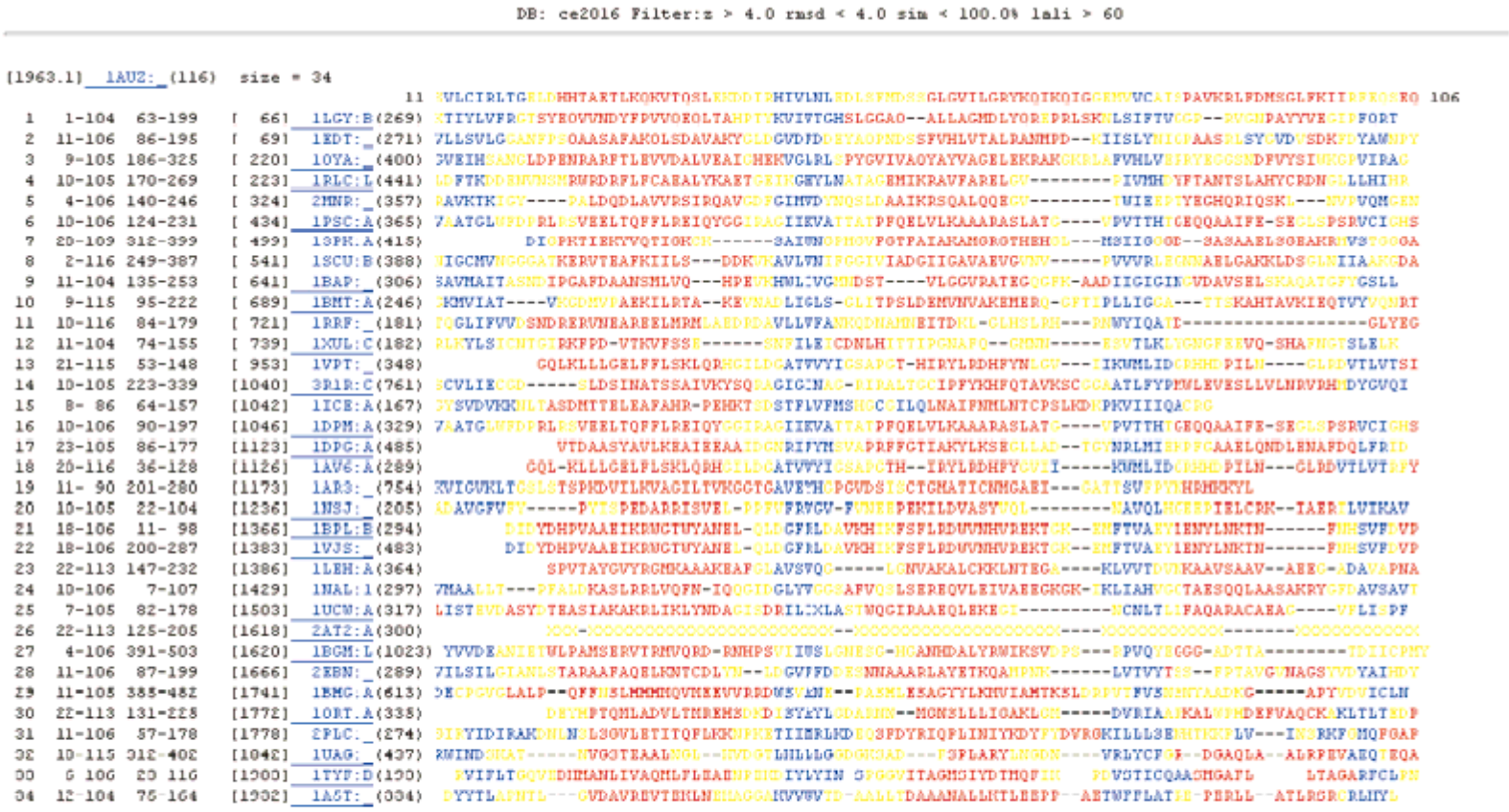
Spinocerebellar ataxia types 2 (SCA2) and 3 (SCA3) are autosomal-dominantly inherited, neurodegenerative diseases caused by CAG repeat expansions in the coding regions of the genes encoding ataxin-2 and ataxin-3, respectively. To provide a rationale for further functional experiments, we explored the protein architectures of ataxin-2 and ataxin-3. Using structure-based multiple sequence alignments of homologous proteins, we investigated domains, sequence motifs, and interaction partners. Our analyses focused on presumably functional amino acids and the construction of tertiary structure models of the RNA-binding Lsm domain of ataxin-2 and the deubiquitinating Josephin domain of ataxin-3. We also speculate about distant evolutionary relationships of ubiquitin-binding UIM, GAT, UBA and CUE domains and helical ANTH and UBX domain extensions.

PMID: 15265035 [PubMed - indexed for MEDLINE]

Example 3 - 20 most frequent common subdomains



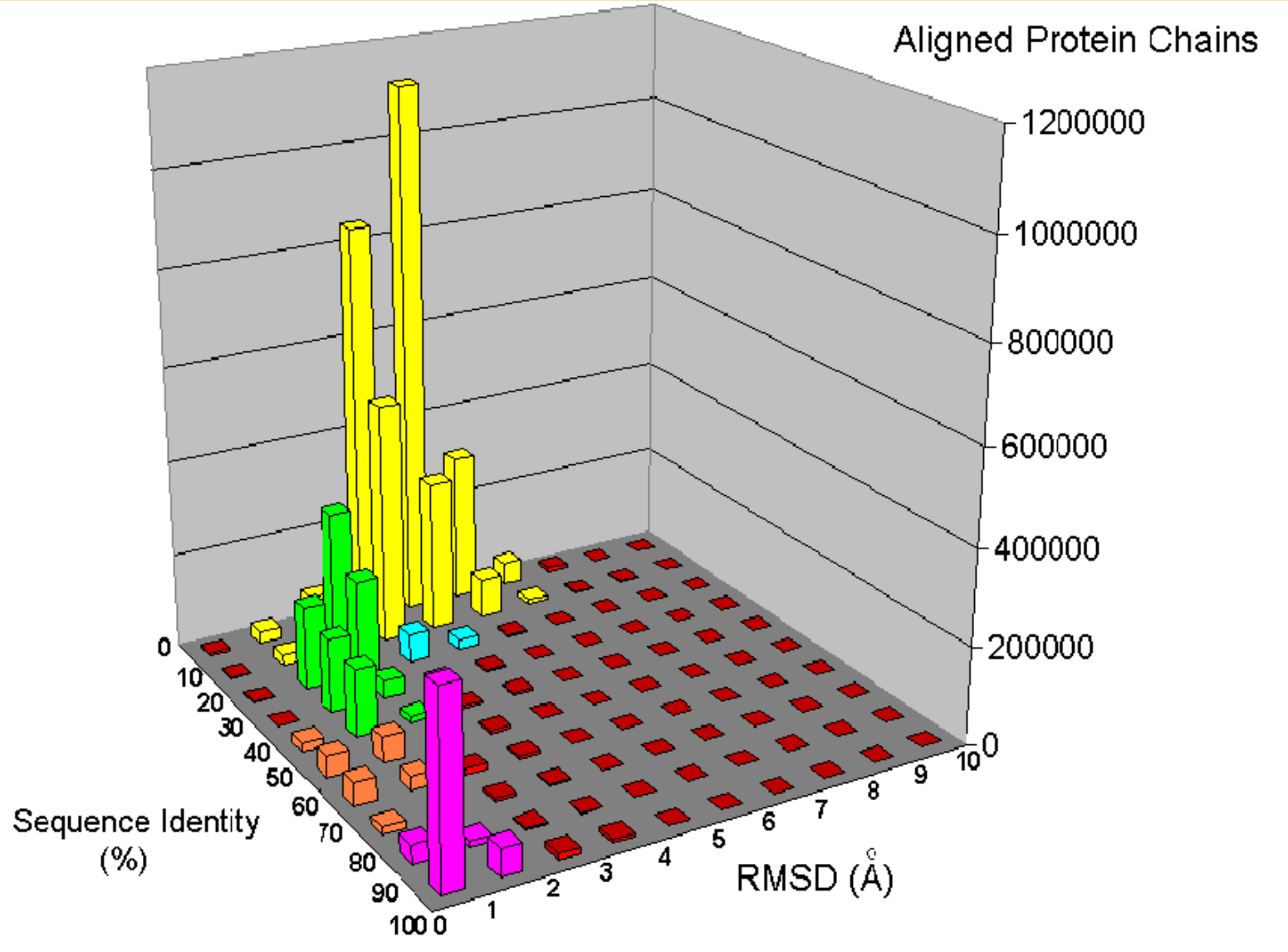
Example 3 – cont.



Detailed view of alignments for substructure [10], spoiiaa, a phosphorylatable component of the system that regulates transcription factor sigmaf of bacillus subtilis (1AUZ:_, residues 11–106) showing specific region of the alignment. The columns represented are: arbitrary reference index; start and end residue number of subdomain structure; start and end residue number of neighbor; index for neighbor in the database; overall length of neighbor chain; and resulting sequence alignment. The sequence is color-coded such that red is alpha helix, blue is beta strand, and yellow is unassigned.

Example 3 – cont.

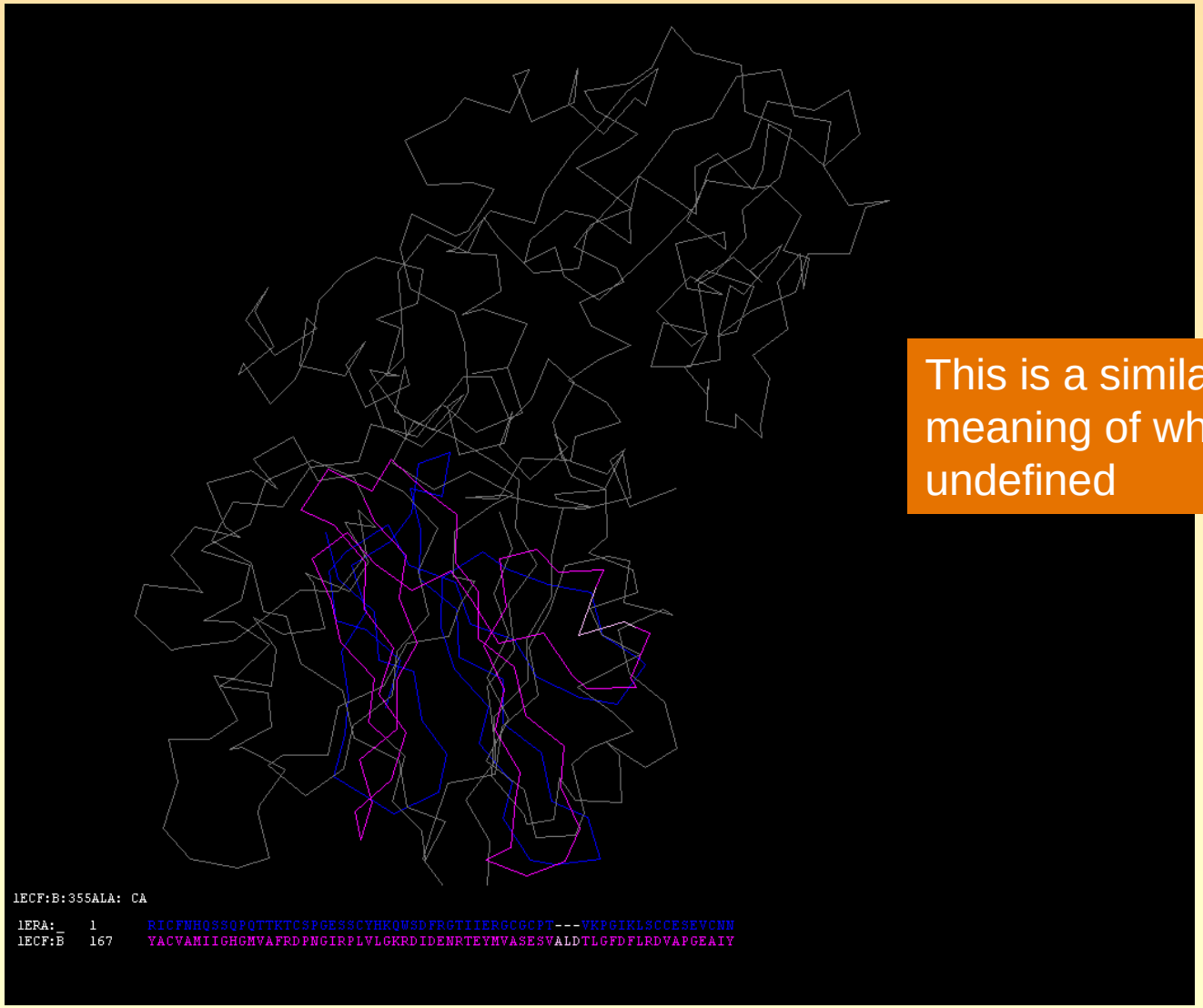
Distribution of structural similarities in PDB detected by CE



Example 4

ERABUTOXIN B (NMR, MINIMIZED AVERAGE STRUCTURE)

GLUTAMINE PHOSPHORIBOSYLPYROPHOSPHATE AMIDOTRANSFERASE



This is a similarity the biological meaning of which still remains undefined

The Future
(also a general rule)

- Gold standards are important
- For structure comparison a human generated alignment standard is important
- Algorithms are then challenged to meet the standard
- Eventually those algorithms highlight problems with the standard
- The cycle continues