

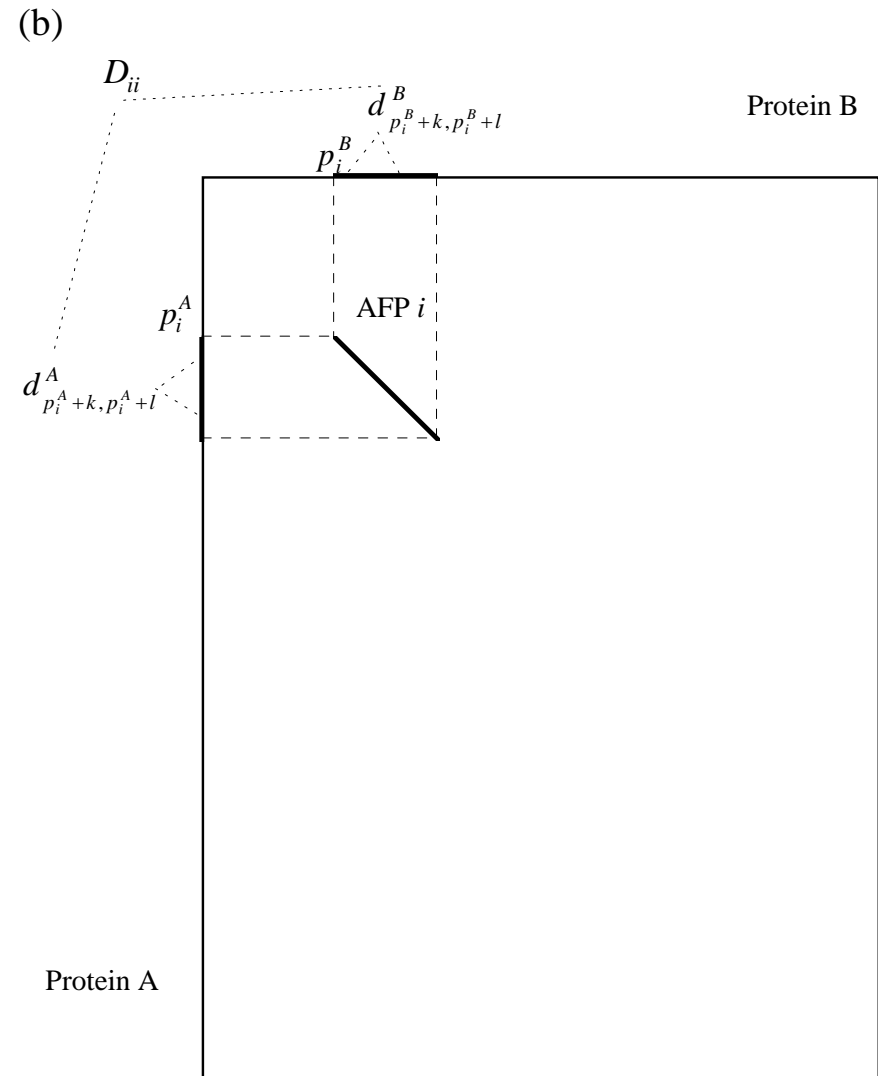
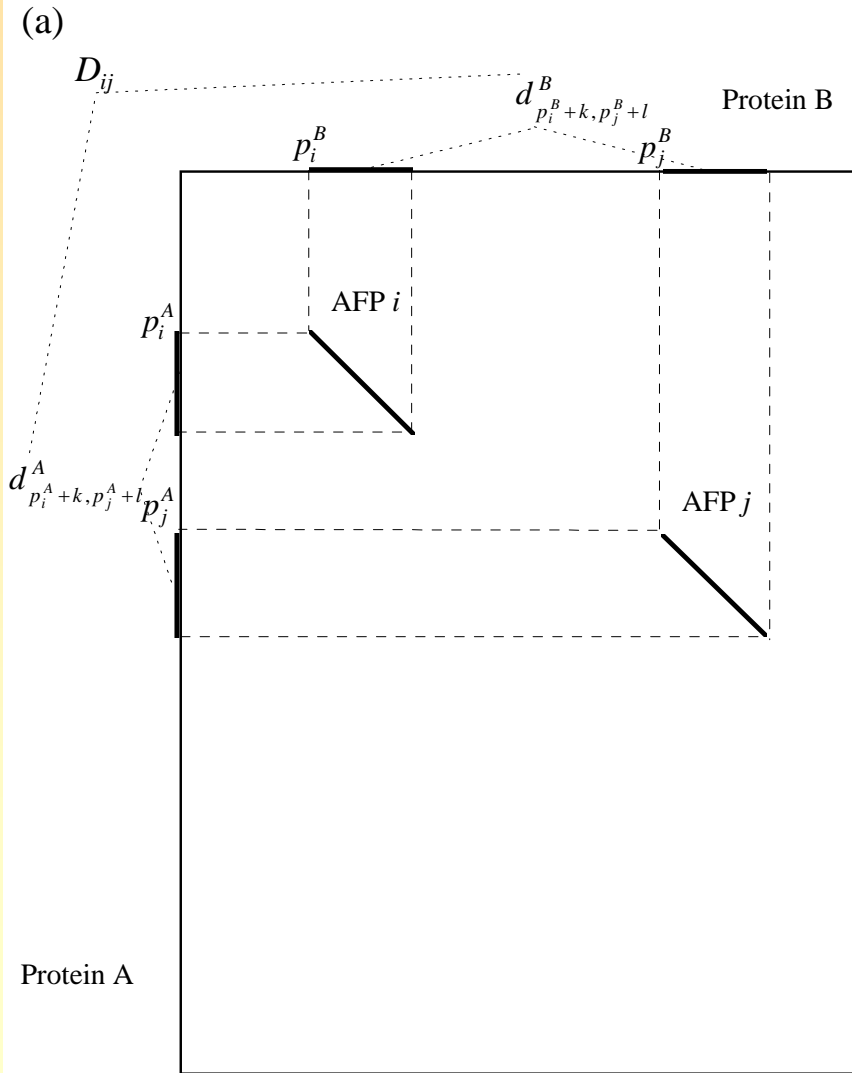
# Understand one method (CE) in more detail

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?

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

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

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

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

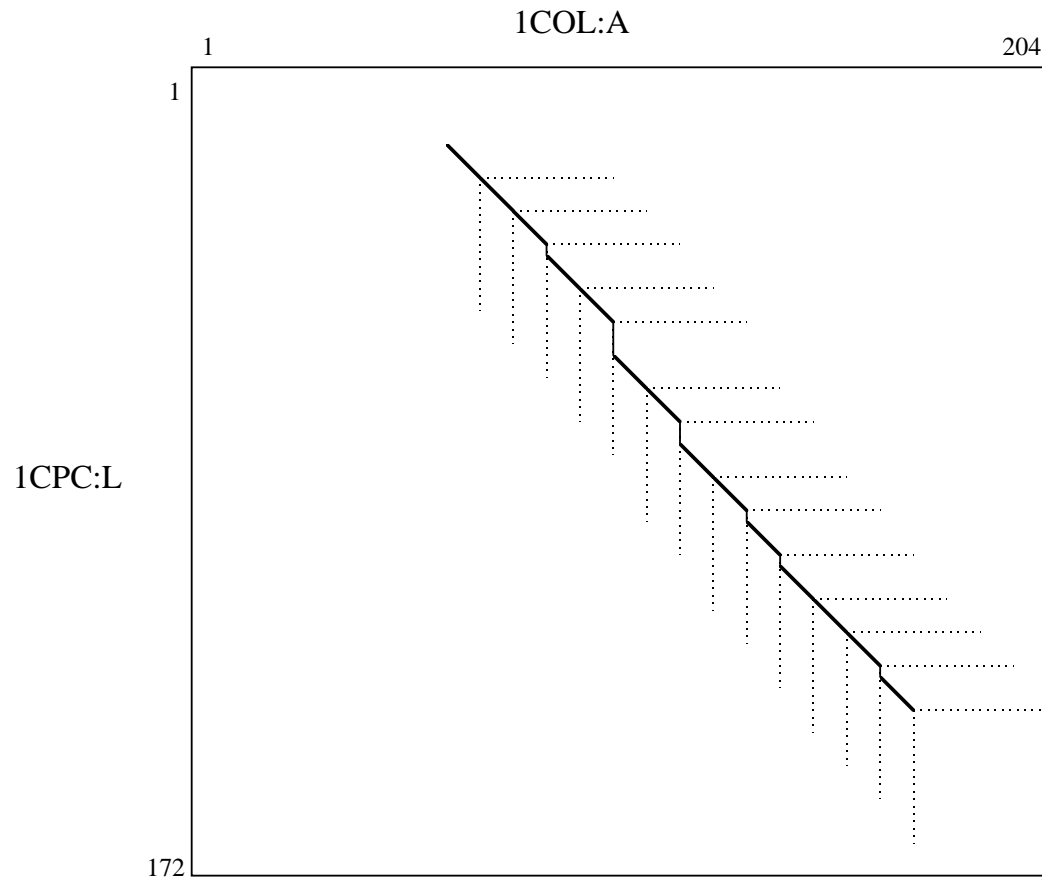
# Optimization of the Final Path

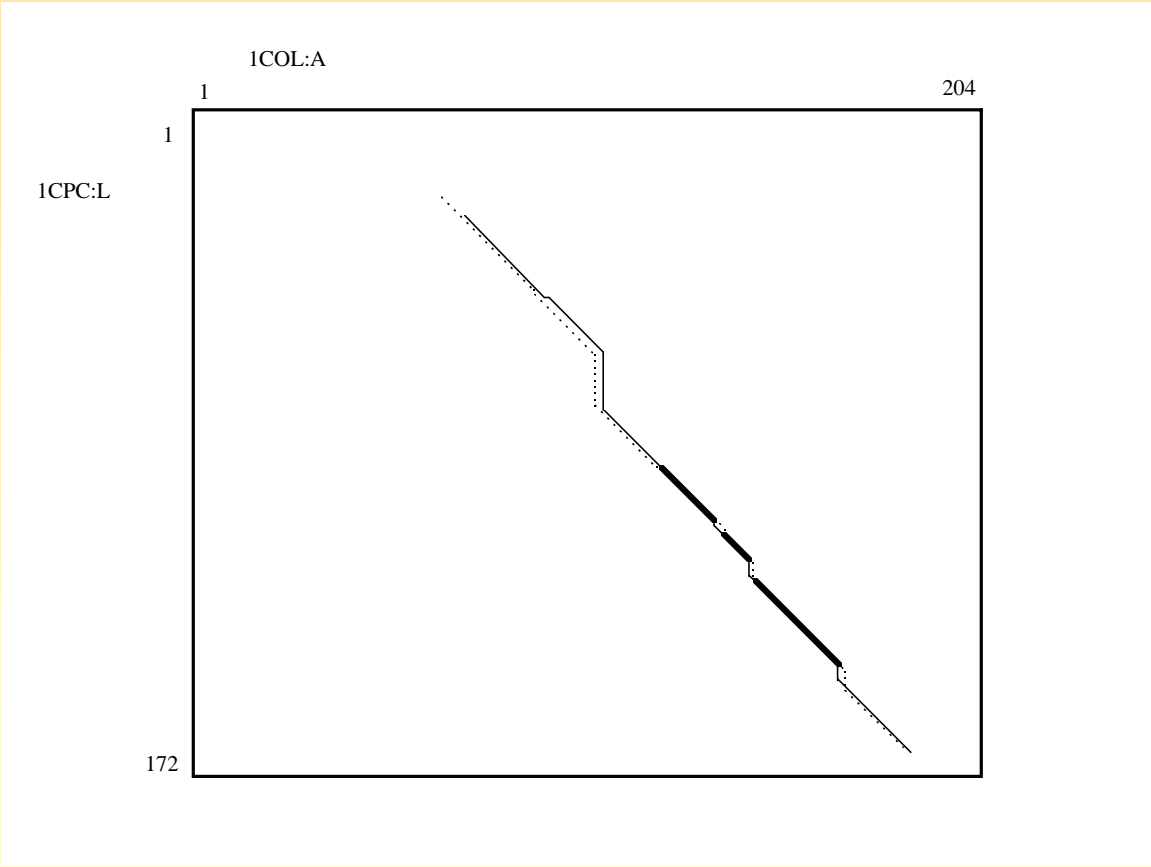
The 20 best alignments with a  $Z$  score above 3.5 are assessed based on rmsd and the best kept. This produces approx. one error in 1000 structures

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

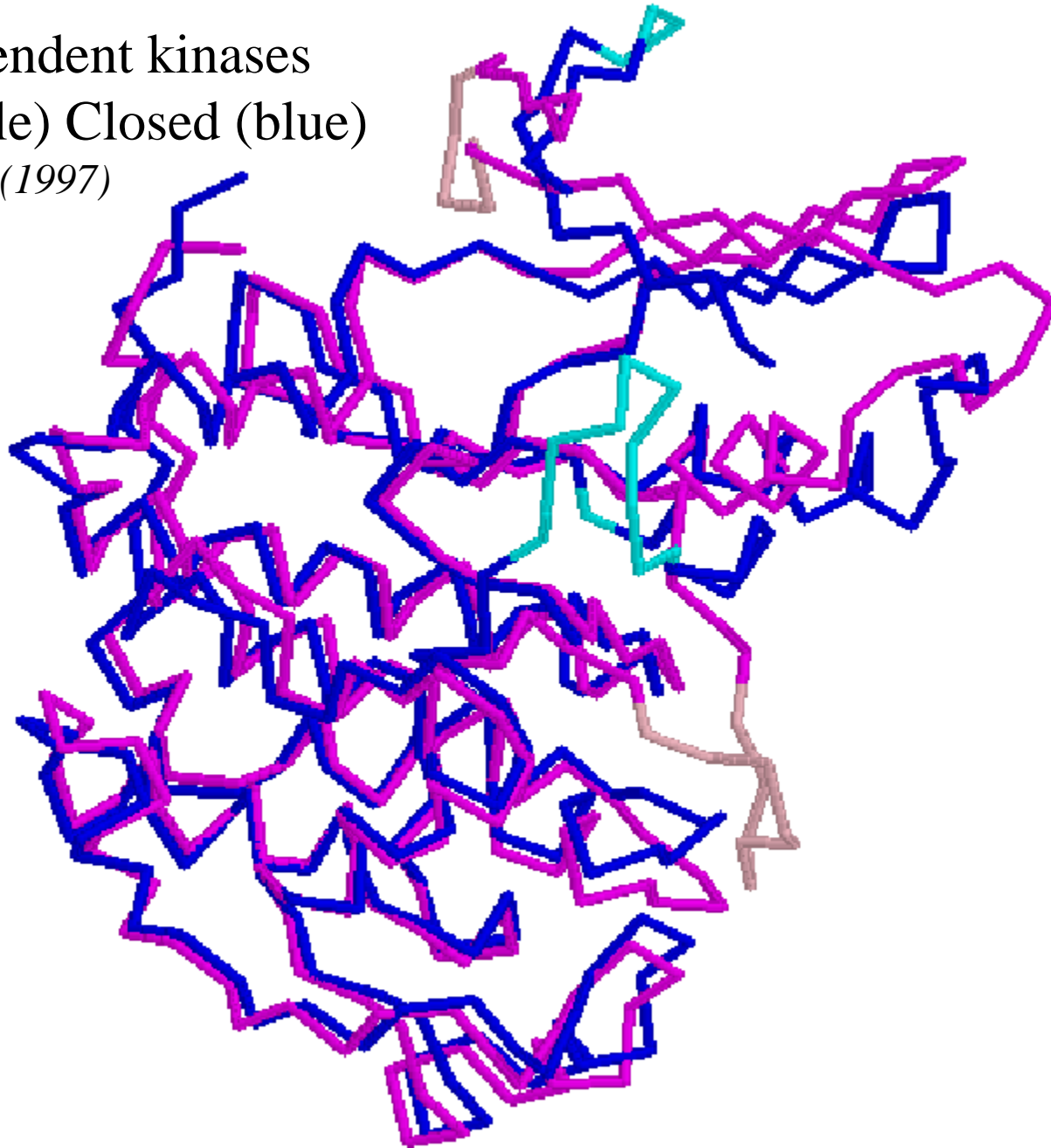
Iterative optimization using dynamic programming is performed using residues for the superimposed structures

# Test Case: Phycocyanin versus Colicin A

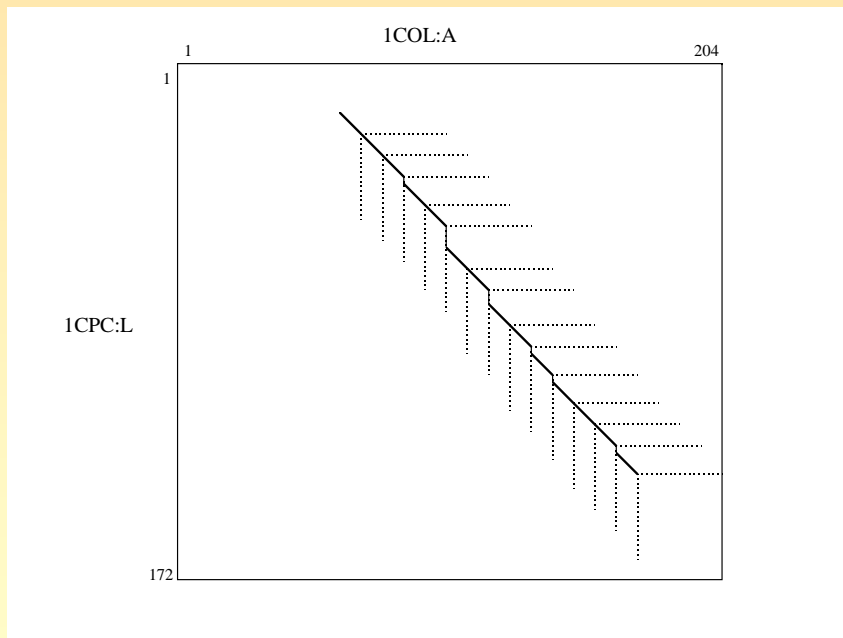




Cyclin-dependent kinases  
Open (purple) Closed (blue)  
*Pavelitch et al. (1997)*



# 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
- $45,000^2/2 * 30 \text{ seconds} = 963 \text{ yrs}$
- 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 [\[PDF\]](#).

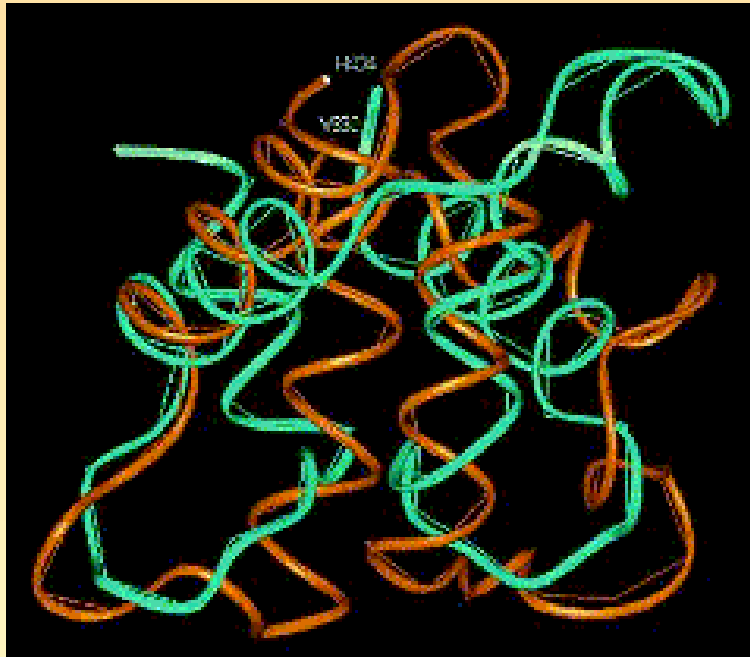


# One Criterion for Redundancy

- 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

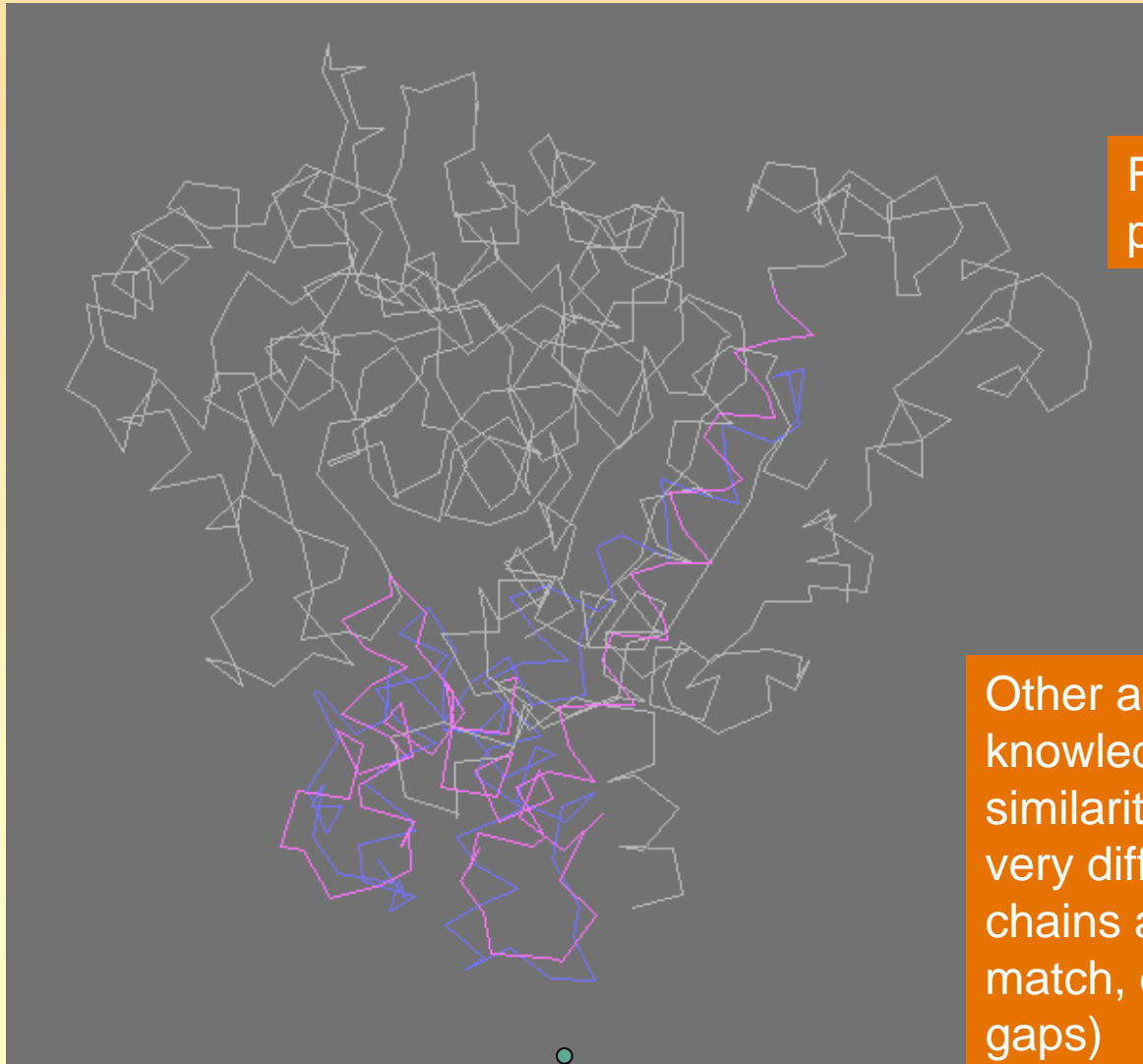


Structural similarity between Acetylcholinesterase and Calmodulin found using CE (Tsigelny et al, *Prot Sci*, 2000, **9**:180)

- CE revealed putative  $\text{Ca}^{++}$  binding domain in acetylcholine esterase
- Sequence similarity to neuroligins predicts  $\text{Ca}^{++}$  binding too – confirmed experimentally
- Members of the a/b hydrolase family bind  $\text{Ca}^{++}$  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

The screenshot shows the PubMed website interface. At the top, there are logos for NCBI, PubMed, and the National Library of Medicine (NLM). Below the logos is a navigation bar with tabs for Entrez, PubMed, Nucleotide, Protein, Genome, Structure, OMIM, PMC, Journals, and Books. A search bar is present with the text 'PubMed' and a 'Go' button. Below the search bar are options for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. A 'Display' dropdown is set to 'Abstract', and 'Show:' is set to '20'. There is also a 'Send to' dropdown set to 'Text'. On the left side, there is a blue sidebar with links for 'About Entrez', 'Text Version', 'Entrez PubMed', 'PubMed Services', and 'Related Resources'. The main content area shows a search result for '1: Eur J Biochem. 2004 Aug;271(15):3155-70.' with a 'Full text' link from Blackwell-Synergy. The title of the paper is 'Structural and functional analysis of ataxin-2 and ataxin-3.' by Albrecht M, Golatta M, Wullner U, and Lengauer T. The authors' affiliation is the Max-Planck-Institute for Informatics, Saarbrücken, Germany. The abstract text describes the study of spinocerebellar ataxia types 2 (SCA2) and 3 (SCA3) and the protein architectures of ataxin-2 and ataxin-3. The PMID is 15265035.

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

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.**

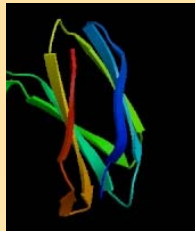
Max-Planck-Institute for Informatics, Saarbrücken, Germany. mario.albrecht@mpi-sb.mpg.de

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.

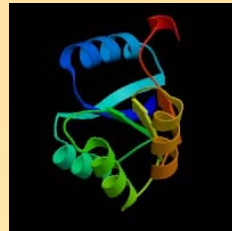
PMID: 15265035 [PubMed - indexed for MEDLINE]

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

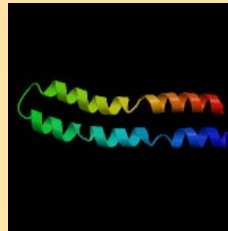
# Example 3 - 20 most frequent common subdomains



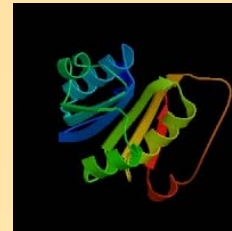
1TEN: \_ 3-89



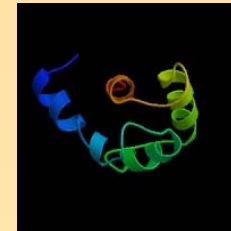
[2] 1RNL: \_ 5-114



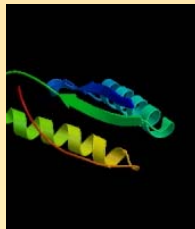
[3] 1A9l: \_ 6-77



[4] 1LDE:C 179-317



[5] 1SMG: \_ 13-86



1TIG: \_ 6-81



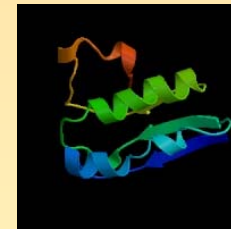
[7] 1PDO: \_ 2-97



[8] 1OPG:A 29-160



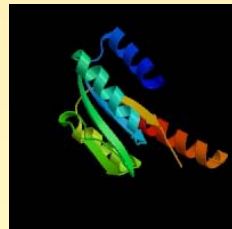
[9] 1AV6:A 47-185



[10] 1AUZ: \_ 11-106



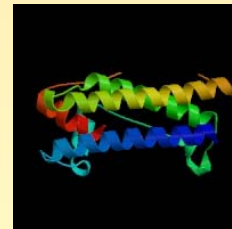
] 2PIA: \_ 100-228



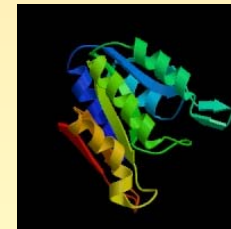
[12] 1VPT: \_ 59-180



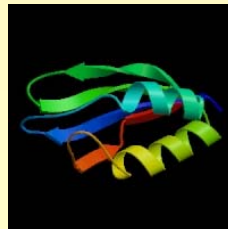
[13] 1IL7: \_ 19-129



[14] 1BGD: \_ 12-154



[15] 1OXP: \_ 104-265



# Example 3 – cont.

DB: ce2016 Filter: z > 4.0 rmsd < 4.0 sim < 100.0% lali > 60

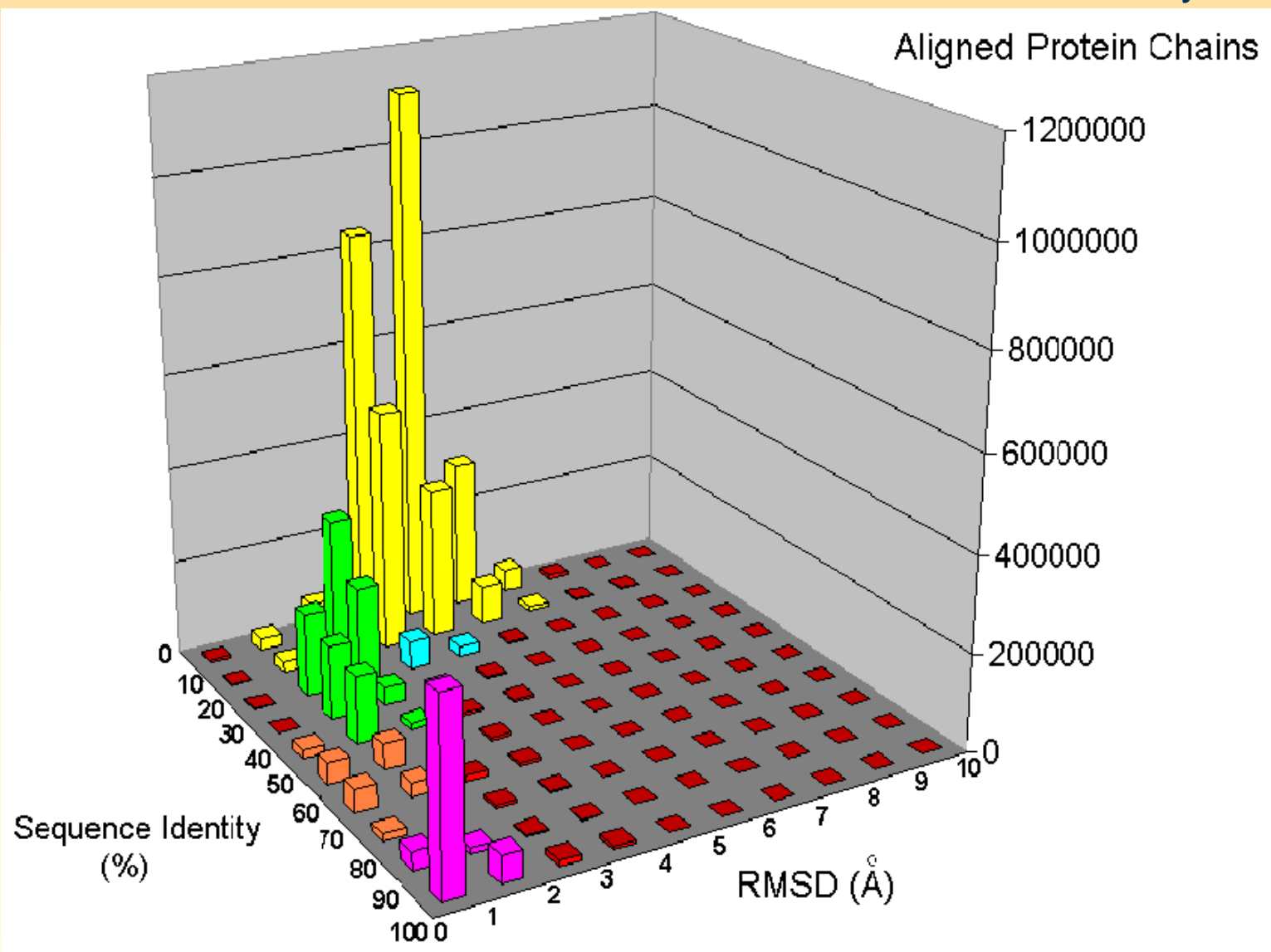
[1963.1]	1AUZ:	(116)	size = 34		
				11	SULCIRLTGLDHHHTARTLKQKVTQSLKDDTPTHTVLNLSDLSFMDSSGLGVTLGRVQTKQIGCEMVCATSPAVKRLFDMSGLFKIIRFQKSRQ 106
1	1-104	63-199	[ 66]	1LCY: B (269)	KTIYLVFRGTSYRQVUNDYFPVUGROLTAHTPTKVIVTGHSLGCA--ALLAGHDLYORRRLSKNLSIFTVGCP--DVCNDAYVVEGIPFORT
2	11-106	86-195	[ 69]	1EDT: _ (271)	VLLSVLGCANFPPOAASAFAROLSDAVARTGLDQVDFFDDEYAOPNDSSFVHLVLTALPAMHFD--KIISLYNICFAASRLSTGVQVDSKDFYAMNPFY
3	9-105	186-325	[ 220]	1OYA: _ (400)	DVEIHGANGLDPENRARTLEVVLDLVEAICHKEKVLRLSPYGVVIAOYATVAGELEKRAKCKRBLAVHVLVRIEYEGCSNDPVYSIHKQVPIRAC
4	10-105	170-269	[ 223]	1RLC: L (441)	LDPTKDDHNVNHRWRDRFLPCABALYKAEITGRIINGEYINATAGEHIKRAVFARELGV-----PIVMDYPTANTSLAHYCRDNGLLLHHR
5	4-106	140-246	[ 324]	2MNR: _ (357)	RAVKTKICY----PALDQDLAVVRSIQAVGDFGIIHVDYNSLDAAIKRSQALQOQGV-----TWIEPTTECHQRIQSKL---NVPVQMGEN
6	10-106	124-231	[ 434]	1PSC: A (365)	YAATGLWRDPRLESVEELTQFFLREIQYGGIRAGIEVATTTATPPQELVLKAAARASLATG---VPVTHIQEQAAIFE--SEGLSPSRVCI GHS
7	20-109	312-399	[ 499]	1SPK: A (415)	DIQPKTIEKTVQTIQKCK-----SAIWNPHOVVPTFAIAKAMOROTHEHGL---MSIIGOOD--SASAAELSOBARKRNVSTGOGA
8	2-116	249-387	[ 541]	1SCU: B (388)	NIGCMVHGCGATKERVUTBAFKIILS---DDKVTAVLVNIFGGVIADGII GAVAEVGVV---PVVRLRGNWALCAKRLDSGLNIIAAGCDA
9	11-104	135-253	[ 641]	1BAP: _ (306)	SAVMHITASNDIPGAFDAANSHLVQ---HPEVHNLVCGMHDST---VLGCVRATEGCGPK--AADIIIGICINVDVAVSELKAGATGFYGSLL
10	9-115	95-222	[ 689]	1EHT: A (246)	SKMVIAT---VKGDMVFAEKILRTA---KEVNADLIGLS--CLITPSLDEMNVVAKHEHRQ--CFTIPLLIGCA---TTSKAHTAVKIEQTVVYQVRT
11	10-116	84-179	[ 721]	1RRF: _ (181)	IQGLIPVVDSDNDRRUVNBAEELMRLAEDPDVLLVFAKQGDHAMHBITDKL--CLHSLPH---RWYIQAIT-----GLYEG
12	11-104	74-155	[ 739]	1XUL: C (182)	RLHYLSICHTGIRKFPD-VTKVFSSE-----SNFILEICDNLHITTIQNAFO--GDMN-----ESVTLKLYGNGFEVQ--SHAFTGTSLELK
13	21-115	53-148	[ 953]	1VPT: _ (348)	GQLKLLGLLFFLSKLRHCILDGATVYVYCSADPT--HIRYLRDHFVNLGV---IIRKMLIDGRHDPIILN---CLRDVTLVTSI
14	10-105	223-339	[1040]	3RIR: C (761)	SCVLIKCCD-----SLDSINATSSAIVKYSQKACIGCNAG--RIRALTGCIFFYKHFQTAVKSCCGAATLFFYPMWLVESLLVLRVPHMDYCVQVI
15	8- 86	64-157	[1042]	1ICE: A (167)	DYSDVYKMLIASDMITTELEAFAMR--PEHRTSDTFLVMSHCC--ILQNALFNNLHNTCPSLRKMPHVI IQACRG
16	10-106	90-197	[1046]	1DPM: A (329)	YAATGLWRDPRLESVEELTQFFLREIQYGGIRAGIEVATTTATPPQELVLKAAARASLATG---VPVTHIQEQAAIFE--SEGLSPSRVCI GHS
17	23-105	86-177	[1123]	1DPC: A (485)	VTDAAASYAVLKRAIEEAAIDGRIFFYNSVAPRFFCTIAKYLKSEGLLAD--TCVNRMLMIEPFCAAELQMDLBNAPDQLFRID
18	20-116	36-128	[1126]	1AVG: A (289)	GQL-KLLCELEFLSKLRHCILDGATVYVYCSADPT---IRYLRDHFVNLGV---IIRKMLIDGRHDPIILN---CLRDVTLVTSI
19	11- 90	201-280	[1173]	1ARS: _ (754)	KVICWKLTSLSLSPMDVILRVAGILTVKGGTGAUEYHGPCVDSISCTGHATICNHGAEI---GATTSVFPYRHHGQYLL
20	10-105	22-104	[1236]	1NSJ: _ (205)	ADAVGVVY----PVYSPEDARRISVEL--PFPVFRVGV--FVNEEPERILDVASYVQL-----NAVQLHCEEPILCRK--IARRLVIRAV
21	18-106	11- 98	[1366]	1EPL: E (294)	DIDYDHPVAABIKRNVCTUYANEL--QLDGFELDAVGHINFSFLDGVNHRVREKTK--ENFTVAEYIENYLNKTN-----FMSVFPVVP
22	18-106	200-287	[1383]	1VJS: _ (483)	DIDYDHPVAABIKRNVCTUYANEL--QLDGFELDAVGHINFSFLDGVNHRVREKTK--ENFTVAEYIENYLNKTN-----FMSVFPVVP
23	22-113	147-232	[1386]	1LEH: A (364)	SPVTAYGVYRCHGAAAKBAFGI AVSTQC-----LGNVARALCKKLNTEGA-----KLVVTDVRKAAVSAAV--AHEG--ADAVAPNA
24	10-106	7-107	[1429]	1NAL: L (297)	ZMAALLT---FPALDKASLRRLVQFN--IQGIDGLYVCGSAFVQSLSEREQVLEIVAECKGK--IKLIAHVGCTAESQOLAASAKRYGFI AVSAVT
25	7-105	82-178	[1503]	1UCW: A (317)	LISTEVDASYDTEASIAKAKRLIKLYNDACISDRILXLASTWQCIRAAEQLEKECI-----NENLTLLFAQARACARAC---VFLISPF
26	22-113	125-205	[1618]	2ATZ: A (300)	YOK-XXXXXXXXXXXXXXXXXXXXXXXXXXXX--XXXXXXXXXXXXXXXXXXXXXXXXXXXX--XXXXXXXXXXXXXXXXXXXX
27	4-106	391-503	[1620]	1BGM: L (1023)	YVVDENIETVLPAMSERVTRVQRD--RHNPSVITUSLGNESC--HGAMHDALYRWIKSWDPS---FPVQYEGGC--ADTTA-----DIICPMY
28	11-106	87-199	[1666]	2EEN: _ (289)	VILSILGIANLSTARAFAQELKNTCDLYN--LDGVDFDESNNHAAARLAYETKQANPNK-----LVIVYTS--PPTAVGVNAGSVVDYAIHDY
29	11-105	385-482	[1741]	1EMG: A (613)	DECPVGLALP--QFFNSLMHMMQVYKVEVREDWVKN--FASMLSEACTYLNKVIAMTKSLRPPVTVFVSRNHTAADKG-----APTVDVICLN
30	22-113	131-225	[1772]	1ORT: A (335)	DEYHPTQLADVLTNREMSDMDISTYLLGDANH--MGNSLLLIQAKLUN-----DVRIAAKALWPHDFVVAQCKAKLTLTSDP
31	11-106	57-178	[1778]	2PIC: _ (274)	QIPYIDIRAKDNLNLSQVLETTIQFLRKNPKSTIIRLMDQSPDYRIQFLINIYKYDPTDVRGKILLLSERHTKQFLV---INSKNFMQFGAP
32	10-115	312-402	[1842]	1UAG: _ (437)	KWINDSKAT----NVGGTEAALNGL--HVDGTLHLLLGDDGNSAD---FQFLARYLNGDN---VRLYCFSE--DGAQGLA--ALRPEVANQTEQA
33	6 106	20 116	[1900]	1TYF: D (190)	SVIFLTQVVEDIHANLIVAQMLFLAENPSEIYVYINSPGGVITAGHSIYDTHQFIN--FDVSTICQAACHGATL--LTAGARTCLFN
34	12-104	75-164	[1932]	1AST: _ (394)	DYYTLAPHTL---GVDAVRVETIRLNHRAGGKRVVVVD--AALLTDAAMNALLMHTLEPP--AETWFFLATSE--PERLL--ATLRGRCLMHYL

Detailed view of alignments for substructure [10], spoIIaa, a phosphorylatable component of the system that regulates transcription factor sigma<sup>4</sup> 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

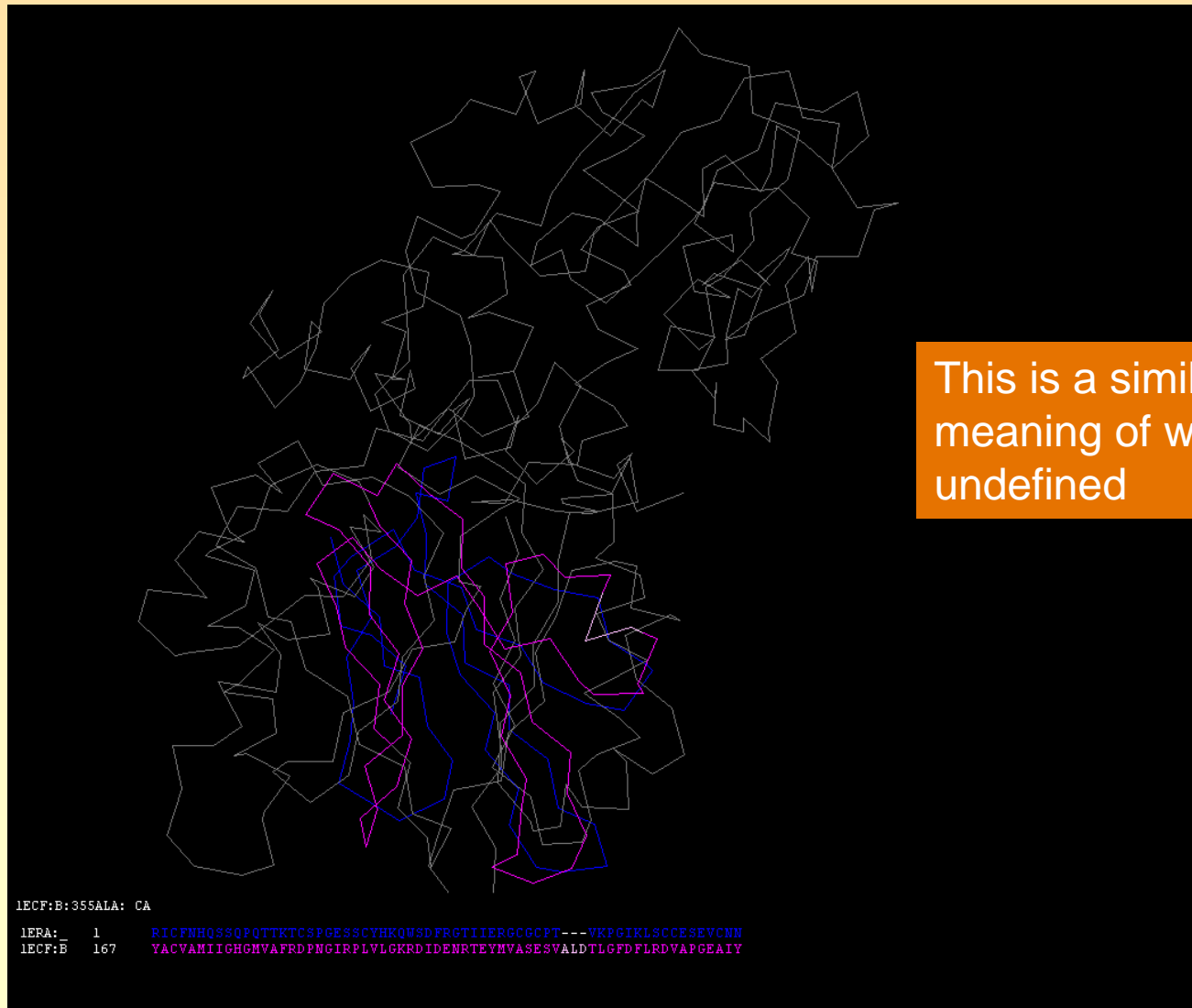




## Example 4

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