Sequence Alignment

Homayoun Valafar

Department of Computer Science and Engineering, USC

Alignment Example

- Align the following sequences:
  - HEAGAWGHEE
  - PAWHEAE
  - Gap penalty of -8, extension penalty of -8.

Global Alignment

- Used Needleman-Wunsch algorithm
- Guarantees global alignment

$$F(i,j) = \max \begin{Bmatrix} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{Bmatrix}$$

		H	E	A	G	A	W	G	H	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9

```
H E A G A W G H _ E
- - P - A W H E A E
```

Needleman-Wunsch Algorithm

- Very useful for global alignment of sequences:

VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED 60
VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED
VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED 60

- Global alignment implies close evolutionary relation.
- What if two sequences are distantly related?
  - A large middle section of a protein is deleted.
- Need to perform local alignment.
  - Smith Waterman Algorithm.

Local Sequence Alignment Algorithm

- Use Smith-Waterman
- Trace-back from the highest score
- Start from the largest score and trace back

$$F(i,j)=\max\begin{Bmatrix}0\\F(i-1,j-1)+s(x_i,y_j)\\F(i-1,j)-d\\F(i,j-1)-d\end{Bmatrix}$$

		H	E	A	G	A	W	G	H	E
	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0
W	0	0	0	0	2	0	20	12	0	0
H	0	10	2	0	0	0	12	18	22	14
E	0	2	16	8	0	0	4	10	18	28
A	0	0	8	21	13	5	0	4	10	20
E	0	0	6	13	18	12	4	0	4	16

Sequence Alignment Tools

- Many many tools on the web
- Most common and universally accepted sequence alignment tools is BLAST (Basic Local Alignment Search Tool)
- One other recommended collection of online tools is EXPASY

Basic Local Alignment Search Tool (BLAST)

- BLAST is an algorithm/tool for comparison of biological sequences
- BLAST applies to DNA, RNA and Protein sequences
- “Basic local alignment search tool” by Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.
- BLAST can be downloaded and easily customized to ones needs
- The algorithm emphasizes speed over sensitivity that is critical considering the number of genomes (GOLD database) that have been sequenced
- Much faster than S-W algorithm but can not guarantee optimality of alignment

Overall Operation of BLAST

- Prunes the entire database based on shorter sequences
- Integrates the much smaller matching sequences, and performs its own sequence alignment on the reduced sequences
- Permits alignment of
  - Protein to Protein (blastp)
  - Nucleotides to Nucleotides (blastn)
  - Protein to Nucleotides
  - Nucleotides to Proteins
- Can query against all known sequences or specialized sequences (known function, known structure)
- PDB has its own implementation (recommended for this class)

Examples

- 1I92:A NA+/H+ EXCHANGE REGULATORY CO-FACTOR mutated by 0.5 45 out of 91.

CAAATGCTTCCTTGTCTTTGTTGGTGTTATAAAGGTCCTAATGTTATTGCTTTTCATTGT
GTTATTTCTAAATGGTATCTTGGTCAATATATTGAAGATGTTGATAAACATTTTCCTGCT
ATGTCTGCTTCTATTATTGCTGGTTATGATTGTTTTGAAGTTAATAATAAAAATGTTGAA
AAAACTACTCATCCTGAAGAAGTTTCTTTTATTCTTGCTGCTCGTAATAATAAACGTATG
CTTCTTTGGGATCCTGAACAAGCTGCTCGTCTT

- 1SF0

AHHHHHHGSK MIKVKVIGRN IEKEIEWREG MKVRDILRAV GFNTESAIAK VNGKVVLEDD

EVKDGDFVEV IPVVSGG

Assessing Structural Similarity

Homayoun Valafar

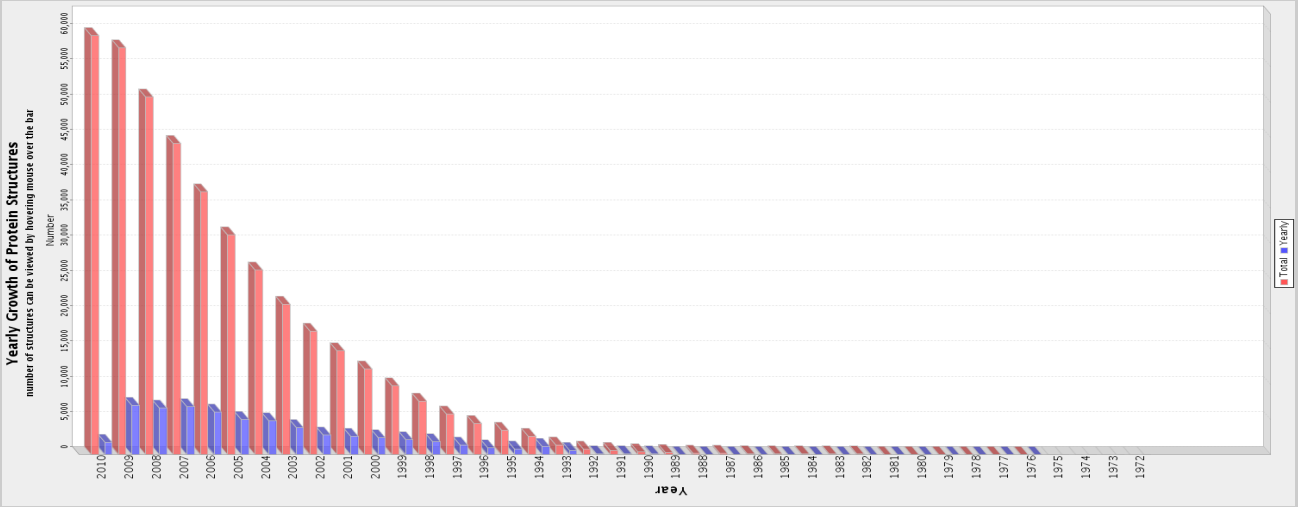
Department of Computer Science and Engineering, USC

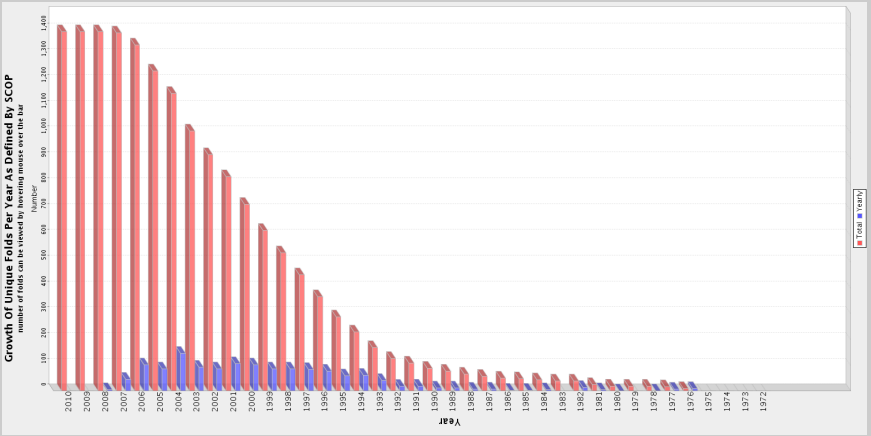
UNIVERSITY OF SOUTH CAROLINA

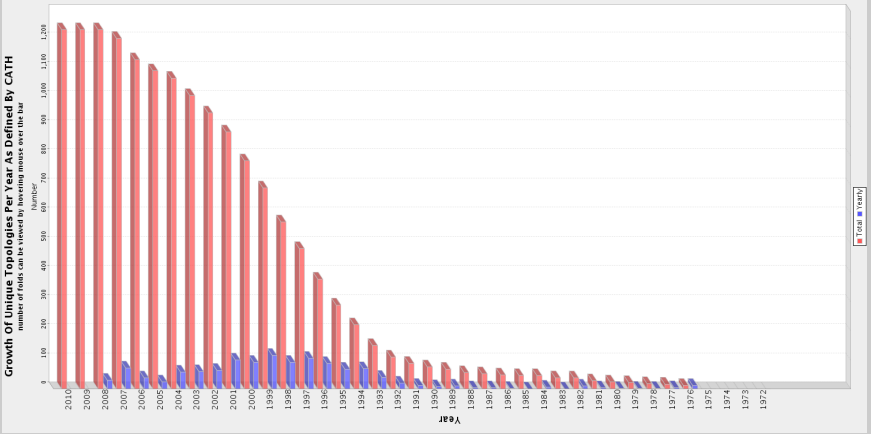
Why Assessing Structural Similarity

- Important in developing structure determination tool for validation
- Important in discovery of functions for unknown proteins (structure leads to function)
- Important in summarizing the core structural information in PDB
  - PDB consists of nearly 60,000 structures but not all of them are unique
  - Search for Lysozyme produced 1189 hits in the PDB
  - How many unique protein structures are there in the PDB?

Status of PDB



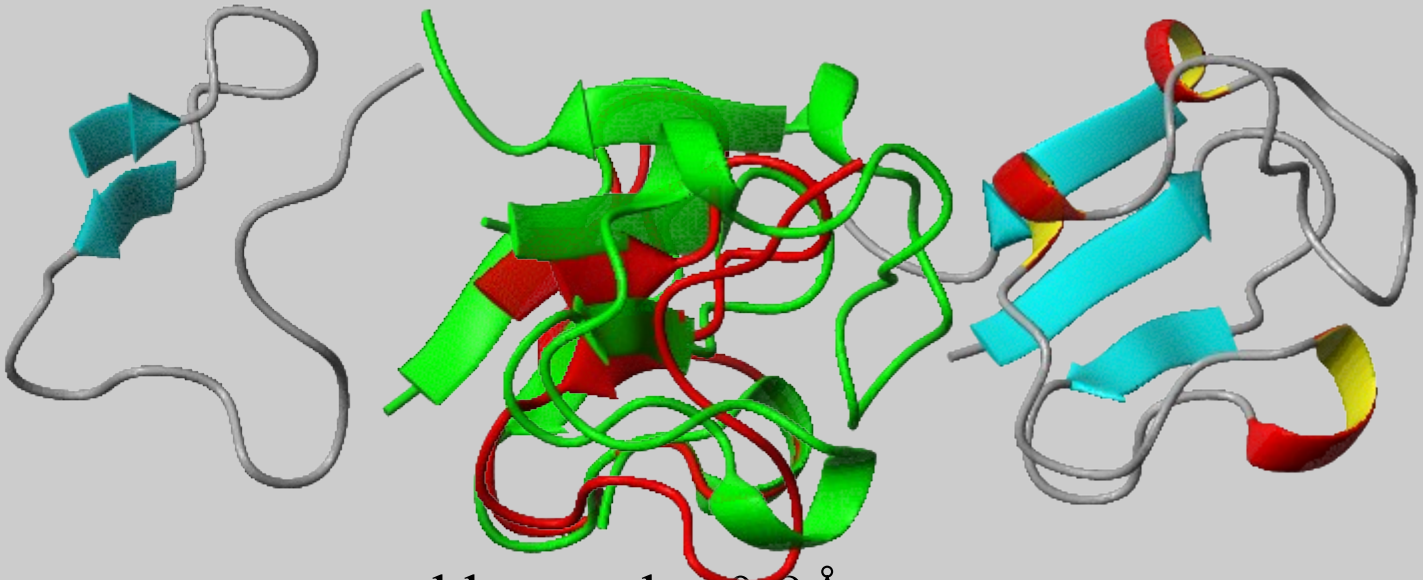




What is Structure Alignment

- Identify and quantify structurally similar regions of two given proteins
- Backbone RMDS is the most commonly used method
  - Assumes a-priori knowledge of portions of the two proteins that need to be matched
    - Example: #1:10-25 to #2:45-60
  - RMSD score may not be fully descriptive of similarity/dissimilarity
    - Low RMSD score is conclusive
    - High RMSD score is not fully conclusive

Examples

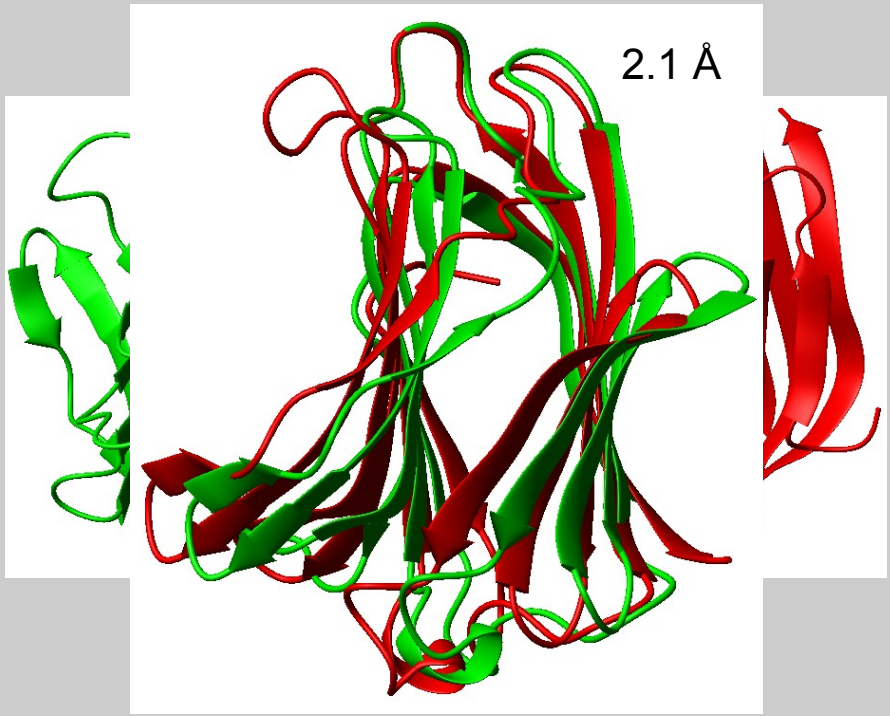


bb_rmsd = 0.8Å

bb_rmsd = 5.8Å

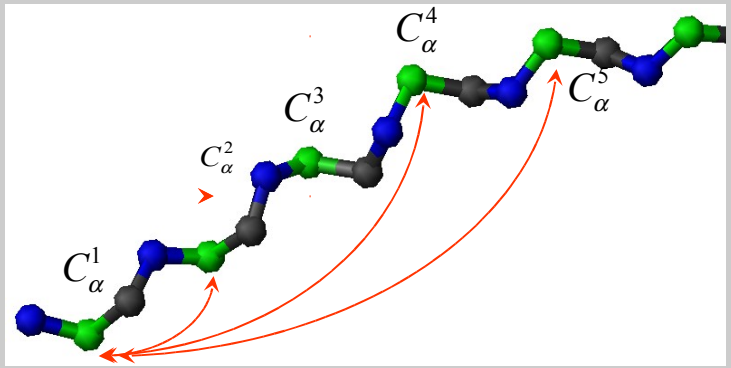
Structure Alignment

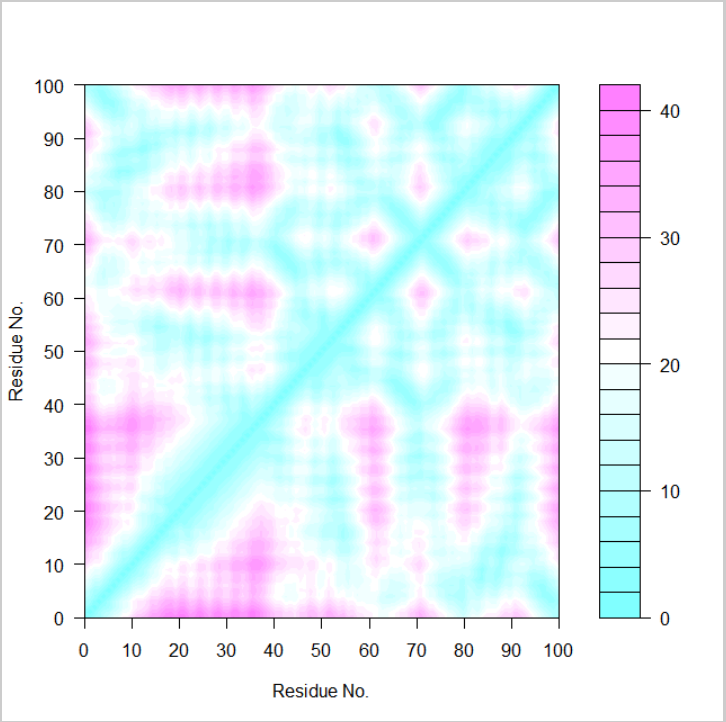
- The number of characterized protein structures is increasing precipitously.
- Analysis of such vast amounts of data requires automated and fast methods.
- A number of structure alignment methods have been introduced over the past few years.
  - DALI
  - SSM
  - CE
  - MAMMOTH
  - TALI (presented here)



Protein Structure Representation

- Most methods operate as follows:
  1. Secondary Structure composition
  2. Connectivity (HBH versus HHB)
  3. Orientation of SSE
  4. Distance matrices (contact maps)
- Distance matrices report pair wise Cartesian distances between all backbone C_α atoms.

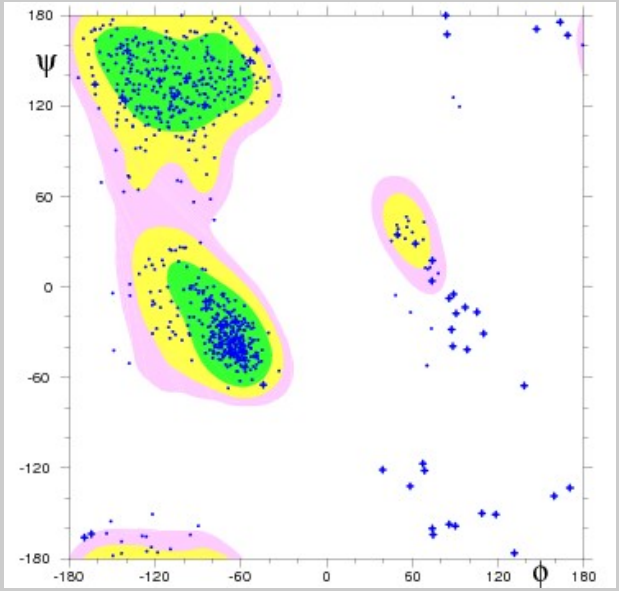




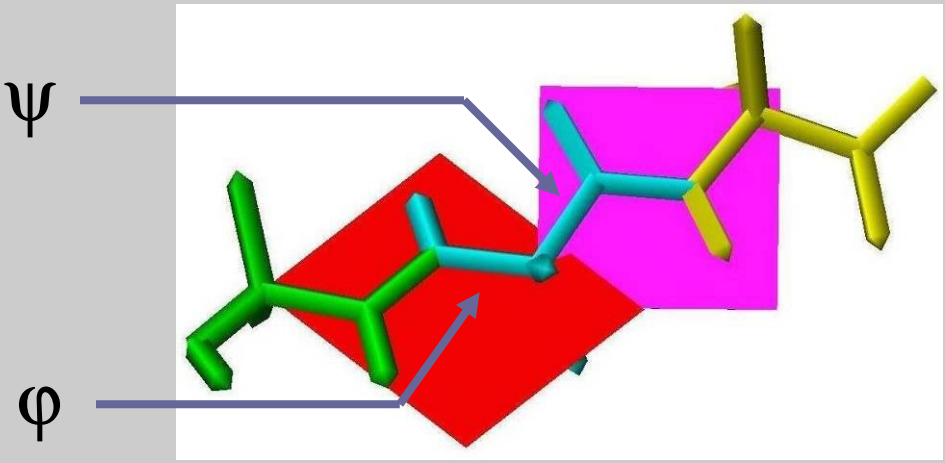
Distance Matrix of first 100 residues of 1CHM

Protein Structure Representation

- TALI utilizes backbone torsion angles.
- Any given structure can be represented as a 2-tuple of torsion angles.



Ramachandran Plot of 1CHM



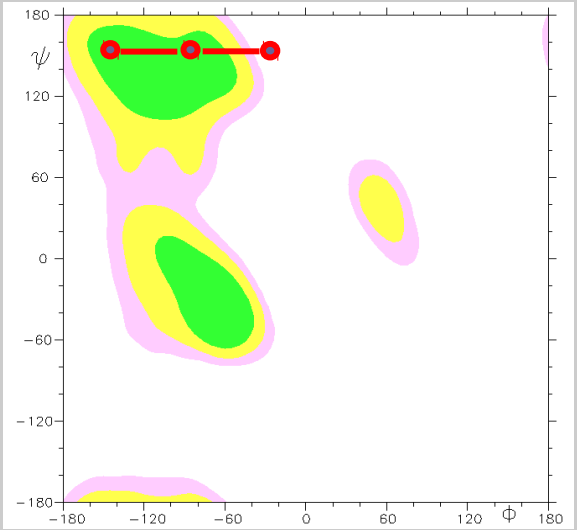
Structure Alignment Using Torsion Angles

- Two structures can be converted to 2-tuple of angles:
  - Structure 1: (-60,-40) (-60,-50) (-100, 0) …
  - Structure 2: (-60,-40) (-60,-40) (-60, -50) (-100,10) …
- Structural alignment can be reformulated as sequence alignment.
  - Can use dynamic programming for alignment.
- A meaningful measure of distance is needed.
  - How to measure the distance between two pairs of angles?
- Can use a number of analytical measures of distance.

Protein Structure Alignment: Torsion angle alignment

- Score Function 1:
  - Euclidian distance between (ϕ_i, ψ_i) and (ϕ_j, ψ_j)

$$d_{ij}[(\varphi_{a,i},\psi_{a,i}),(\varphi_{b,j},\psi_{b,j})] = \sqrt{(\varphi_{a,i}-\varphi_{b,j})^2+(\psi_{a,i}-\psi_{b,j})^2}$$

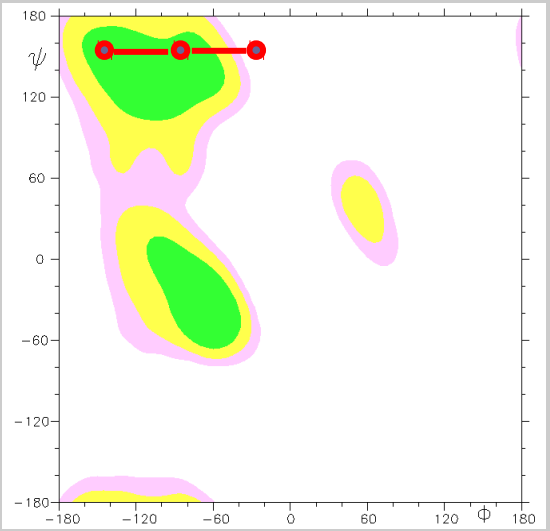


Protein Structure Alignment: Torsion angle alignment (cont.)

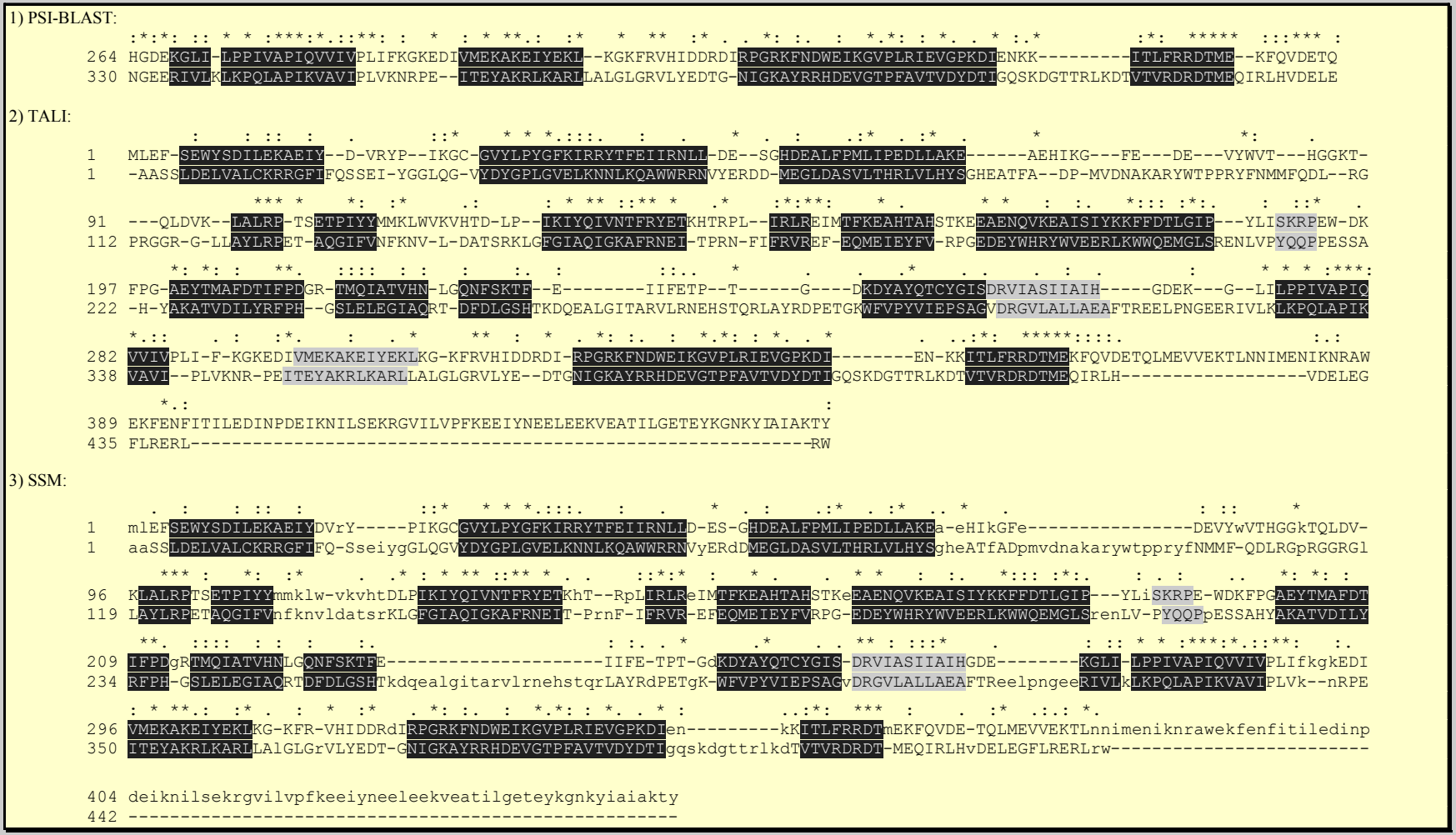
- Score Function 2:
  - Convert the density plot of Ramachandran space to –log().
  - Distance between (ϕ_i, ψ_i) and (ϕ_j, ψ_j) computed by:

$$d_{ij}^{r} = \int_{L} R(l)dl$$

where L is the path connecting two points in space.



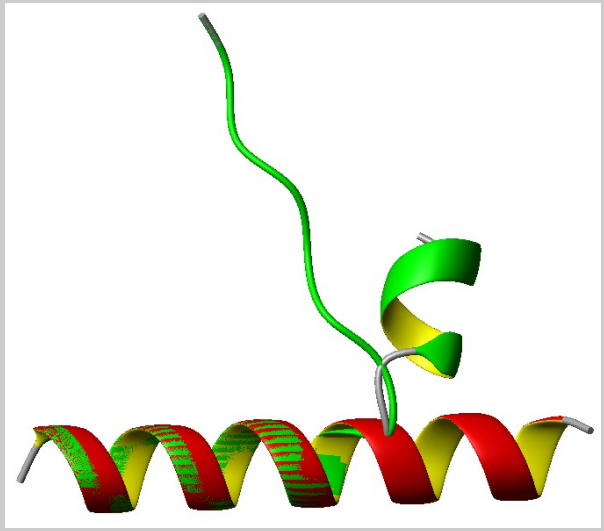
Alignment Results



Fundamental Difference Between TALI and Others

•Other programs identify a set of regions in two structures such that:

$$\left\{\left(R_1^a, R_1^b\right), \left(R_i, R_i\right)\right\}, RMSD\left(\left\{R_1^a, R_1^b\right\}\right) \text{ is minimized}, RMSD\left(\left(R_1^a, R_1^b\right)\right) = 0$$

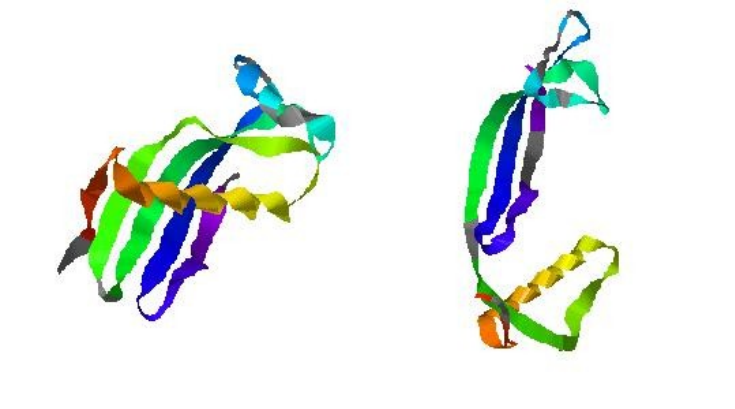


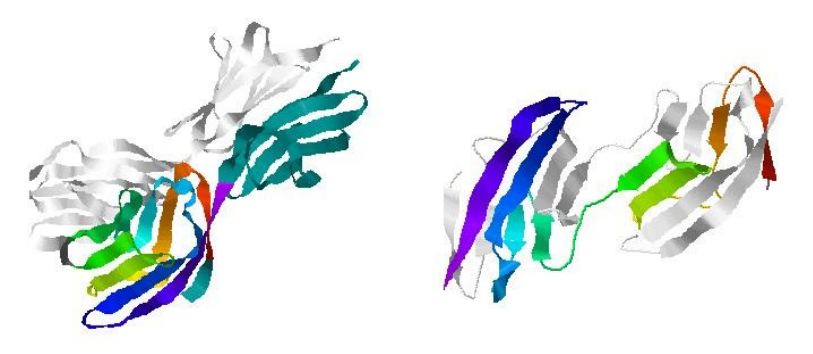
•TALI identifies a set of regions in two structures such that:

$$\left\{\left(R_1^a, R_1^b\right), \left(R_i, R_i\right)\right\}, \sum_i RMSD\left(R_i\right) \text{ is minimized}, RMSD\left(\left(R_1^a, R_1^b\right)\right) + RMSD\left(\left(R_3^a, R_3^b\right)\right) = 0$$

•TALI identifies regions of deviation.

Alignment Result 2





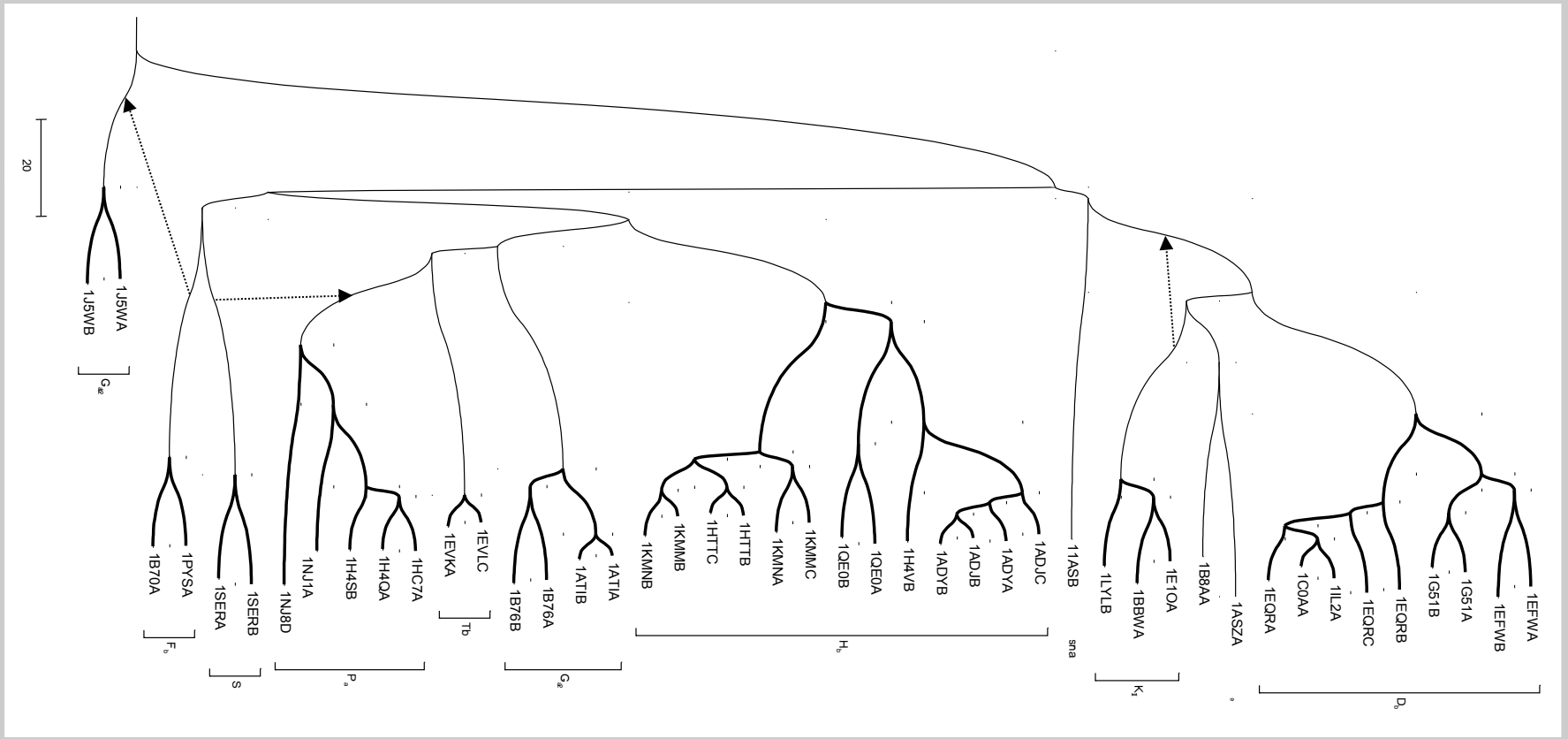
2) 1cewi-1r4ca

```
           : .. ::**::***:**:.***:**** * **.::*: *::*:     ::*:*:**:*****.*: .     .     :  *      : : *:* :*::** . :.*     ...
  1   GAP--VPVDENDEGLQRALQFAMAEYNRASNDKYSSRVVRVISAKRQLVSG--IKYILQVEIGRTTCPKS-S-GDLQSC----E-FHDEPEMAKYTTCTFVVYSIPWLNQIKLL---ESKCQ
  1   --GGPMDASVEEEGVRRALDFAVGEYNKASNDMYHSRALQVVRARKQI--VAGVNYFLDVELGRTTCTKTQPNL---DNCPFHDQPHL-K---RKAFCSFQIYAVPWQGTMTLSKSTCQDA-
```

3) 1hngb-1a64a

```
    ************************************  ****************************************:********
  1 -TVWGALGHGINLNIPNFQMTDDIDEVRWERGSTLVAEFKRKMKPFLKSGAFEILANGDLKIKNLTRDDSGTYNVTVYSTNGTRILNKALDLRIL-
  1 GTVWGALGHGINLNIPNFQMTDDIDEVRWERGSTLVAEFKR--KPFLKSGAFEILANGDLKIKNLTRDDSGTYNVTVYSTNGTRILDKALDLRILE
```

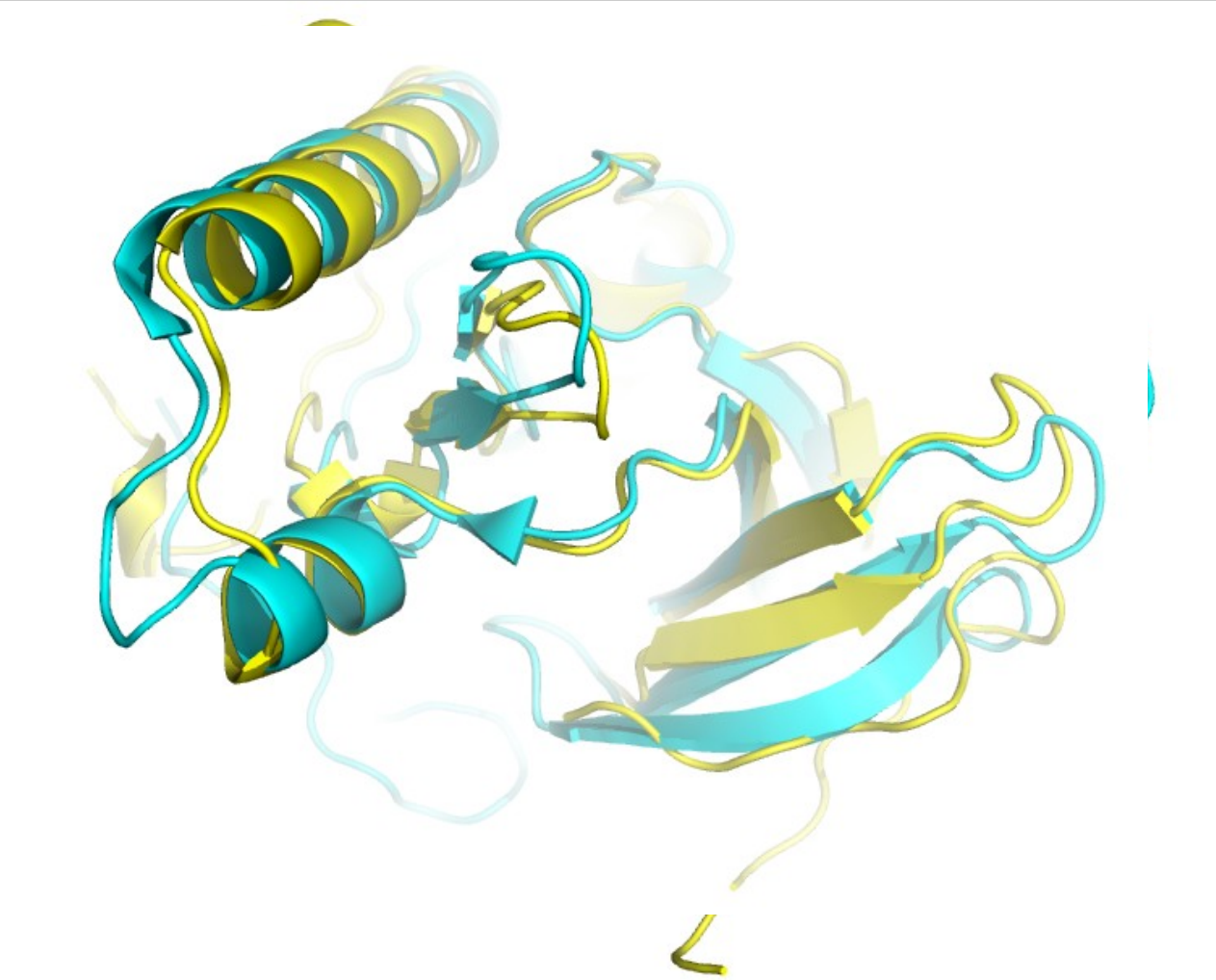
Alignment Result 3: Phylogeny of class II amonoacyl-tRNA synthetase



Why Stop at Torsion Angles

- Residue-to-residue match score:
  - Torsion angles
  - Hydrophobicity
  - Surface accessibility
  - Sequence
  - Distance from protein core
  - Others?

Protein S/R Kinase



Why Not Multiple Structure Alignment?

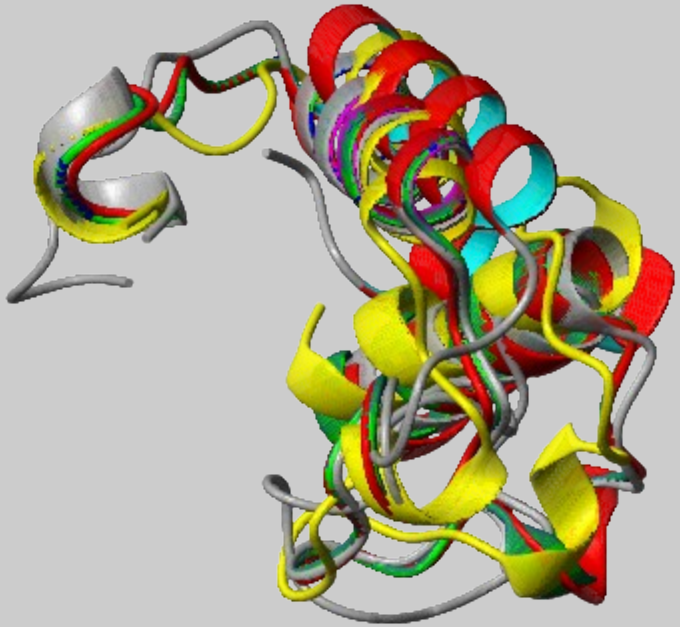
?

Acyl Carrier Proteins

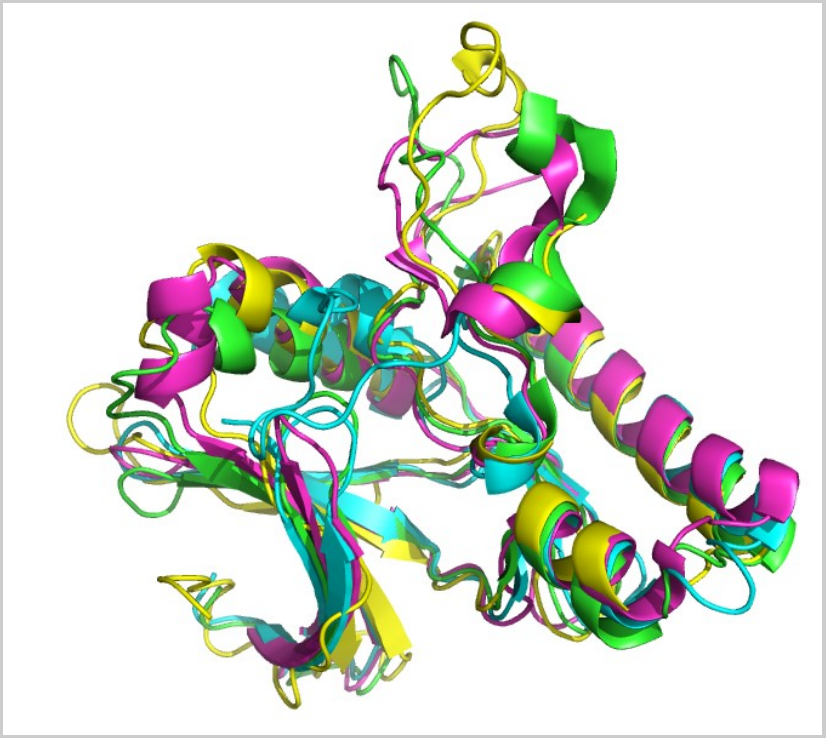
```
2FAC      tieervkkiigeqlgv--kqeevtnnasfvedlgadsldtvelvmaleeefdteipdeeae
1L0I      tieervkkiigeqlgv--kqeevtnnasfvedlgadsldtvelvmaleeefdteipdeeae
AcpP      diaervkkividhlgv--dadkvvesasfiddlgadsldtvelvmafeeefgveipddaad
2JQ4      --natireilakfgqlptpvdtiadeadl-yaaglssfasvqlmlgieeafdiefpdnlln
1ACP      tieervkkiigeqlgv--kqeevtnnasfvedlgadsldtvelvmaleeefdteipdeeae
AcpXL     atfdkvadiiaetsei--dratitpeshtiddlgidsldfldivfaidkefgikiplekwt
2EHS      -leervkeiiaeqlgv--ekekitpeakfvedlgadsldvvelimafeeefgieipdedae

2FAC      lHHHHHHHHHHHHHll--lHHHlllllBlllllllllHHHHHHHHHHHHHHlllllHHHHH
1L0I      lHHHHHHHHHHHHHll--lHHHlllllBlllllllllHHHHHHHHHHHHHHlllllHHHHl
AcpP      lHHHHHHHHHHHHHll--lHHHlllllBlllllllllHHHHHHHHHHHHHHlllllHHHHl
2JQ4      --HHHHHHHHHHlllllllHHHllllllH-HHHlllHHHHHHHHHHHHHHlllllHHHHl
1ACP      llHHHHHHHHHHHlll--lllllllllllllllllllHHHHHHHHHHHHHHlllllHHHHl
AcpXL     lHHHHHHHHHHHHHll--lHHHlllllBlllllllllHHHHHHHHHHHHHHlllllHHHHl
2EHS      -HHHHHHHHHHHHHll--lHHHlllllBlllllllllHHHHHHHHHHHHHHlllllHHHHH
Scores    45889988898888900878898898786888889988999989898878998898888
```

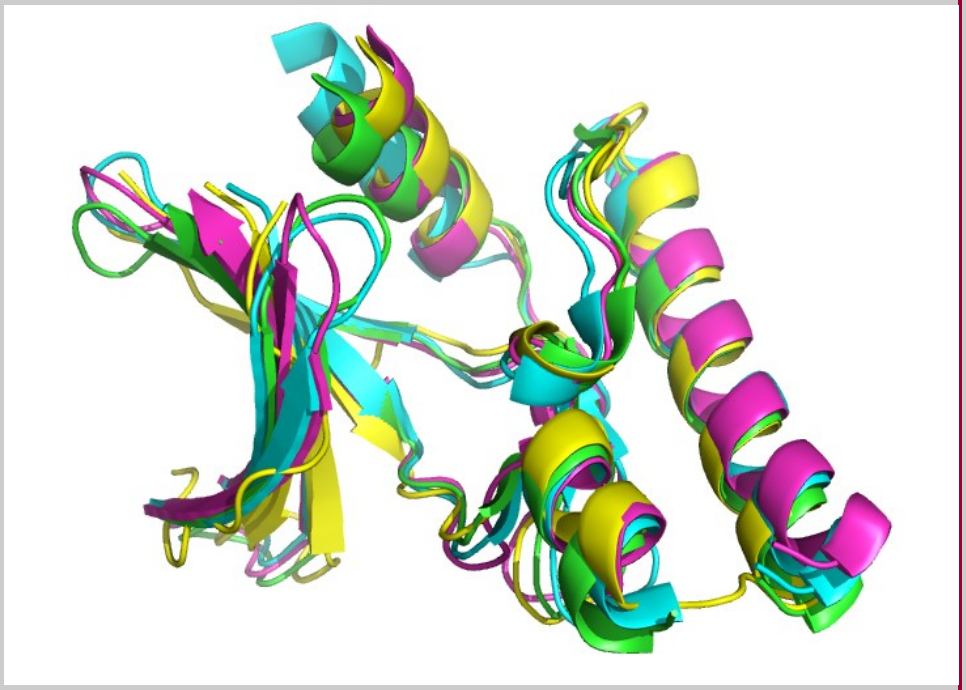
Acyl Carrier Proteins

```
2FAC      ek--ittvqaaidyin---g-hq-
1L0I      ek--mttvqaaidyin---g-hq-
AcpP      ds--iltvgdavkfie---k-aq-
2JQ4      nrksfasikaiedtvklildgkea
1ACP      ek--ittvqaaidyin---g-hq-
AcpXL     tq---e-----vn-----------
2EHS      ek--iqtvgdvinylk---e-k--

2FAC      Hl--llBHHHHHHHHH---H-Hl-
1L0I      ll--llBHHHHHHHHH---H-ll-
AcpP      ll--llBHHHHHHHHH---H-ll-
2JQ4      llHHHHlHHHHHHHHHHHHHlHHH
1ACP      ll--llllHHHHHHHH---H-Hl-
AcpXL     ll---l-----lB-----------
2EHS      Hl--llBHHHHHHHHH---H-H--
Scores    88006766666786660060530
```



Protein Kinase





Protein kinase

```
  nucleotide binding *********                * ATP
1PME      qvfdvgprytnlsyigygmvcsaydnvnkvrvaikkisp-f-ehqtycqrtlreikillrf
1O6L      ------dylkllgkgtfgkvilvrekatgryyamkilrkeviiakdevahtvtesrvlqnt
1UNL      ---qkyeklekigegtygtvfkaknretheivalkrvrldd-ddegvpssalreicllkel
1GZ8      ---enfqkvekigegtygvvykarnkltgevvalkkirv--------pstaireisllkel

1PME      llllllllBBBBBBllllllBBBBBBllllllBBBBBBBll-l-llHHHHHHHHHHHHHHHHl
1O6L      ------BBBBBBBBlllBBBBBBBllllllBBBBBBBBBHHHHHHlllHHHHHHHHHHHHHll
1UNL      ---llBBBBBBBBBlllllBBBBBBBllllllBBBBBBBBlll-llllHHHHHHHHHHHHlll
1GZ8      ---llBBBBBBBBBlllllBBBBBBBllllllBBBBBBBll--------lHHHHHHHHHHlll

Scores    000434877888888768798888887877789897976130334378888898889778
```

Protein kinase

```
                                (ATP) ***  * (ATP)
1PME        frheniigindiiraptieqmkdvylvthlmgadlykllktqhlsndhicyflyqilrglk
1O6L        trhpfltalkyafqth--drlcfvmeyang-gelffhlsrervfteerarfygaeivsale
1UNL        lkhknivrlhdvlhsd--kkltlvfefcdqdlkkyfdscngd-ldpeivksflfqllkglg
1GZ8        lnhpnivklldvihte--nklylvfeflhqdlkkfmdasaltgiplpliksylfqllqgla

1PME        llllBllllBBBlllllllllllBBBBBBlllBBHHHHHHHHllllHHHHHHHHHHHHHHHH
1O6L        llllBllBBBBBBll--lBBBBBBBllll-lBHHHHHHHHllllHHHHHHHHHHHHHHHH
1UNL        llllBllBBBBBBll--lBBBBBBBlllBBHHHHHHHllll-llHHHHHHHHHHHHHHHH
1GZ8        llllBllBBBBBBBl--lBBBBBBBlllBBHHHHHHHlllllllHHHHHHHHHHHHHHHH

Scores      889889878788887600788779887887375588887767388787888889898897
```

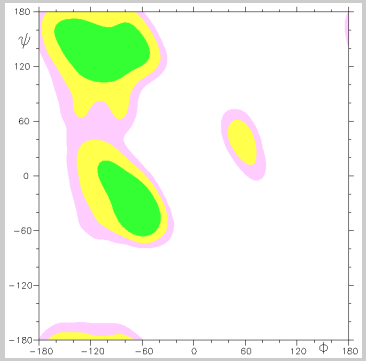
Protein kinase

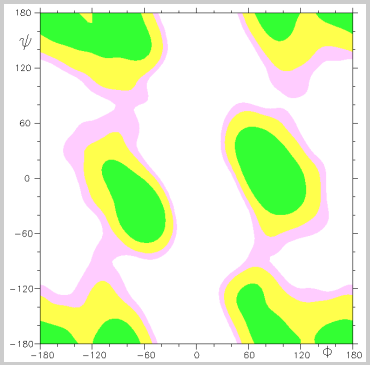
```
     proton acceptor * **** (ATP)      * (ATP)
1PME      kyihsanvlhrdlkpsnlllntt-dlkicdfglarvadpdhdhtg-fl-teyvatrwyr
1O6L      eylhsrdvvyrdiklenlmldkdghikitdfglckegisdgatmk-fc-g--tpeylap
1UNL      gfchsrnvlhrdlkpqnllinrngelklanfglarafgipvrcysaevvtlwyrppdvl
1GZ8      afchshrvlhrdlkpqnllintegaikladfglarafgvpvrtyt-he-v--vtlwyr-

1PME      HHHHHllBBllllllHHHBBBlll-lBBBllllllBBllHHHllll-ll-llllllHHHl
1O6L      HHHHHlllBllllllHHHBBBllllllBBBllllllBlllllllllll-ll-l--lHHHlll
1UNL      HHHHHllBBllllllHHHBBBllllllBBBllllllBBlllllllllllllllHHHllHHHH
1GZ8      HHHHHlllllllllllHHHBBBllllllBBBllllHHHHHllllBlll-ll-B--llllll-

Scores    78899888888988889989887478898889886677677678707707117667773
```

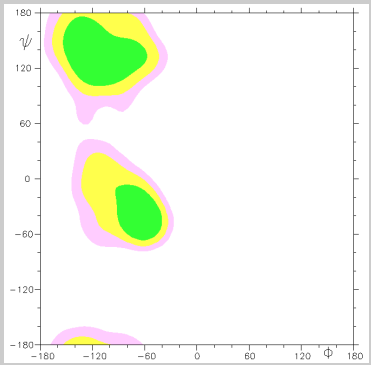
Summary

- Structure alignment can be reformulated as a sequence alignment problem.
- TALI has been demonstrated to be successful in alignment of structures.
- TALI can identify points of structural differences.
- TALI is available online @ ifestos.cse.sc.edu

Future Directions







- TALI can be expanded through the following changes:
  - Integration of amino acid specific Ramachandran space (i.e. GLY, PRO)
  - Simultaneous sequence and structure alignment.
  - Empirically derived scoring matrices.

Acknowledgement

- TALI was developed by Dr. Xijiang Miao
- MsTALI is currently being developed by Paul Shealy