Comparison of Protein Structures

Models, Measures, Metrics and Methods

Natalio Krasnogor
www.cs.nott.ac.uk/~nxk

asap research
automated scheduling optimisation & planning

The University of Nottingham

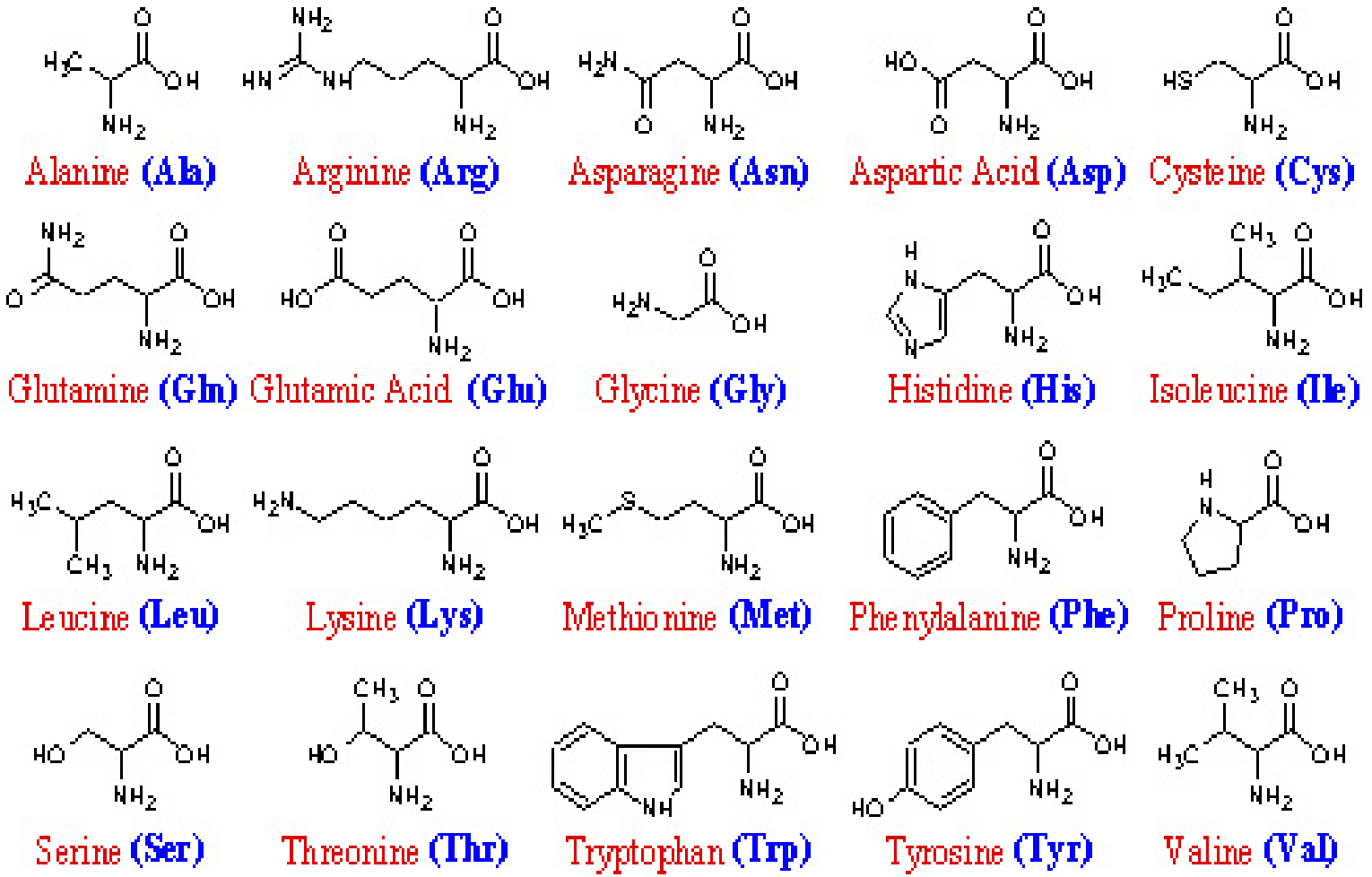
The 3 Minutes Protein Gist

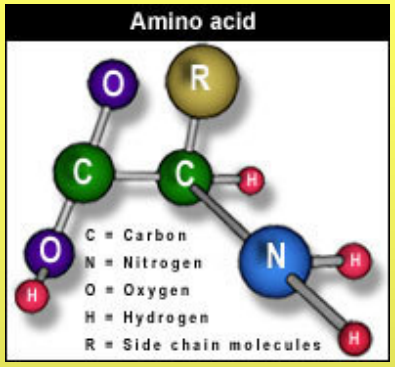
- Proteins are chains of 20 different types of amino acids
- Joined together in any linear order
- This sequence of amino acids is the *primary structure* (represented as a string of 20 different symbols)
- The primary sequence forms *secondary structures*
- The secondary structures form *tertiary structures*

We want to compare these objects!

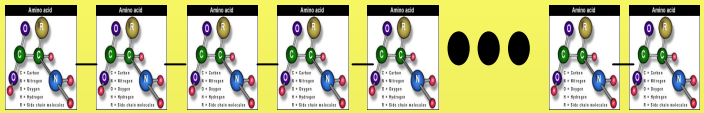
Schematic diagrams of the 20 amino acids

(picture taken from www.chemistry.pomona.edu)

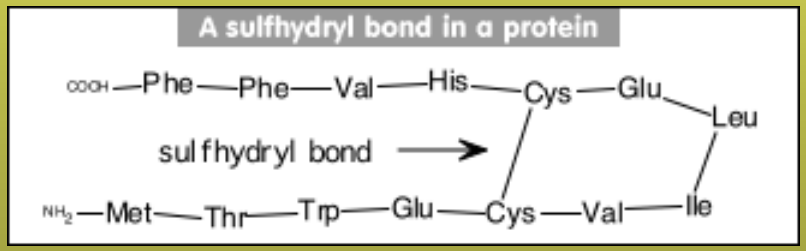




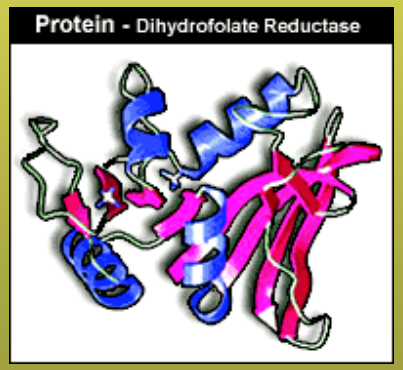
Primary structure



Secondary structure



Tertiary structure



asap research
automated scheduling optimisation & planning

The University of Nottingham

Proteins Role in Life

- **Structural Proteins:** the organism's basic building blocks, eg. collagen, nails, hair, etc.

- **Enzymes:** biological engines which mediate multitude of biochemical reactions. Usually enzymes are very specific and catalyze only a single type of reaction, but they can play a role in more than one pathway.

- **Transmembrane proteins:** they are the cell's housekeepers, eg. By regulating cell volume, extraction and concentration of small molecules from the extracellular environment and generation of ionic gradients essential for muscle and nerve cell function (sodium/potasium pump is an example)

Why do we want to compare tertiary structures ?

- Group proteins by structural similarities
- Determine the impact of individual residues on the protein structure
- Identify distant homologues of protein families
- Predict function of proteins with low degree of primary structure (i.e.. sequence) similarity with other proteins
- Engineer new proteins for specific functions
- Assess ab-initio predictions

Sequence-Structure-Function relationships

1) Conserved 1° sequences ⟶ similar structures

2) Similar structures $\xrightarrow{?}$ conserved 1° sequences

3) Similar structures ⟶ conserved function

Protein engineering

Introduce mutations in genes of an existing protein to alter its STRUCTURE and hence FUNCTION in a pred

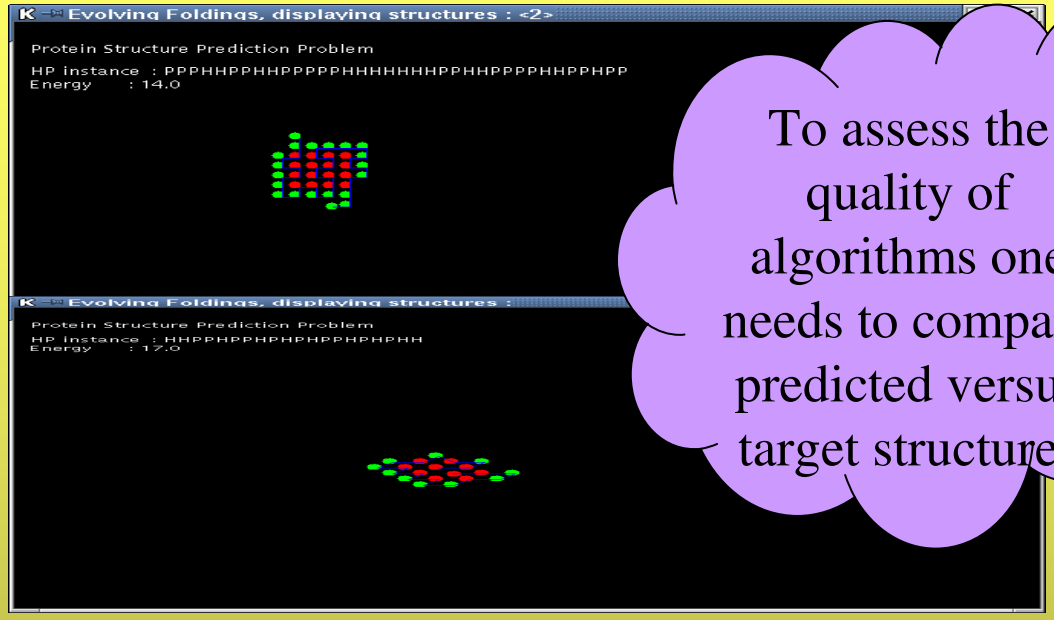
Find out structural changes by comparing with original structure

Example:

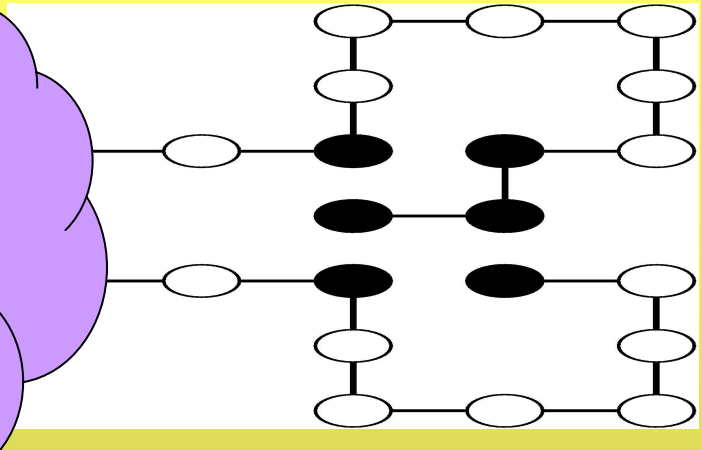
Make a restriction enzyme that ified site in the DNA.

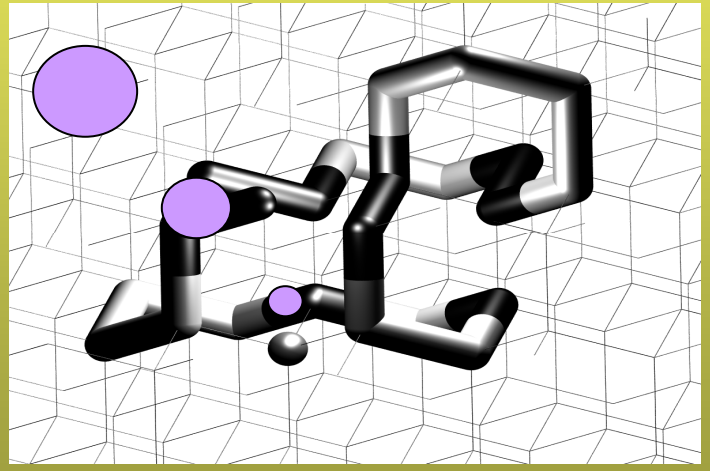
GCATG……TAGCGTATTATTTT

Assessment of *Ab-Initio* Protein Structure Prediction



To assess the quality of algorithms one needs to compare predicted versus target structures

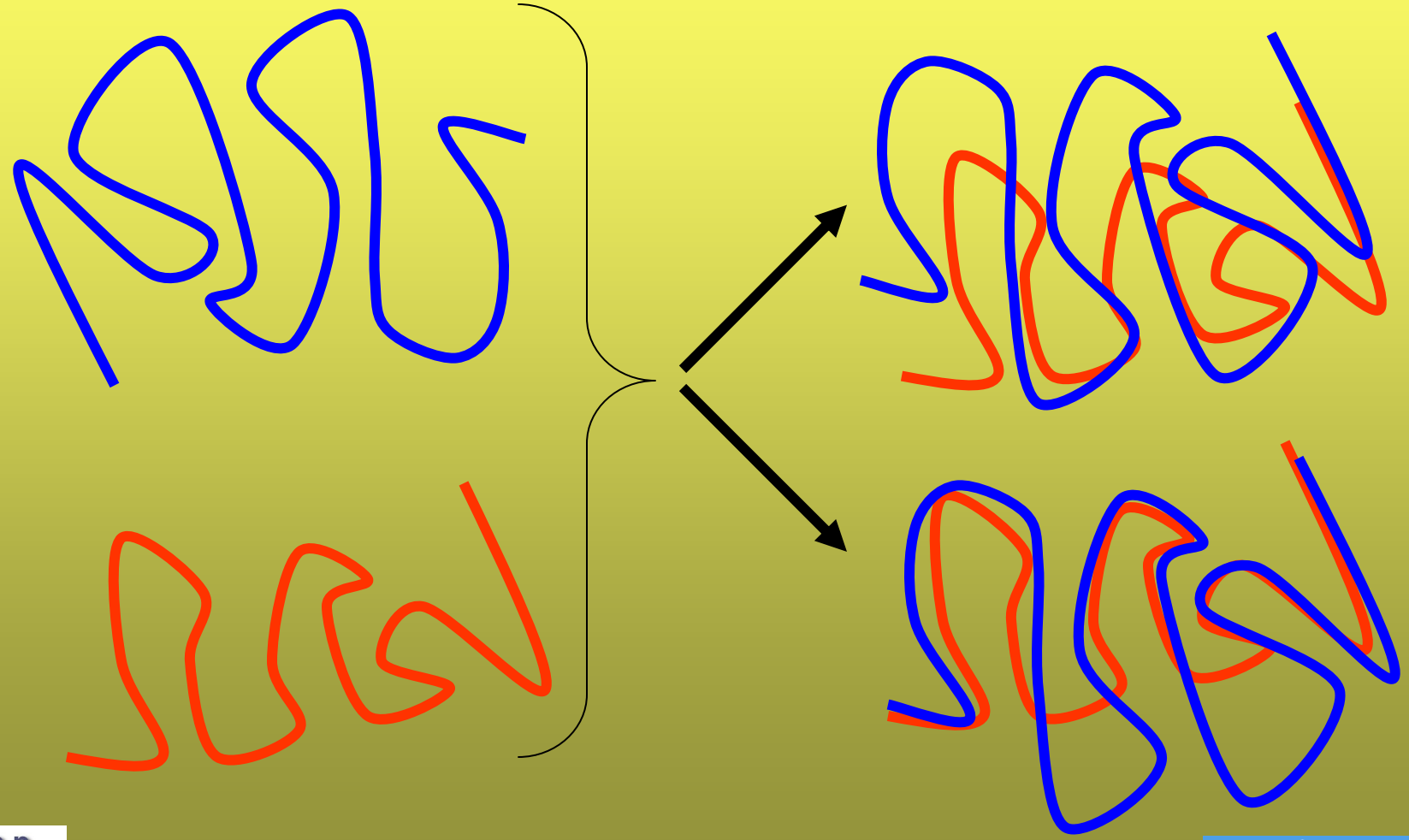




From top left clockwise:

1. Snapshot of optimally solved 2d-square instance
2. Optimal structure for functional model instance (note the non-compact nature of the optimal structure)
3. As 2 but in a diamond (3d) lattice. The sphere shows the binding pocket
4. As 1 but in a triangular lattice.

Comparing Protein Structures



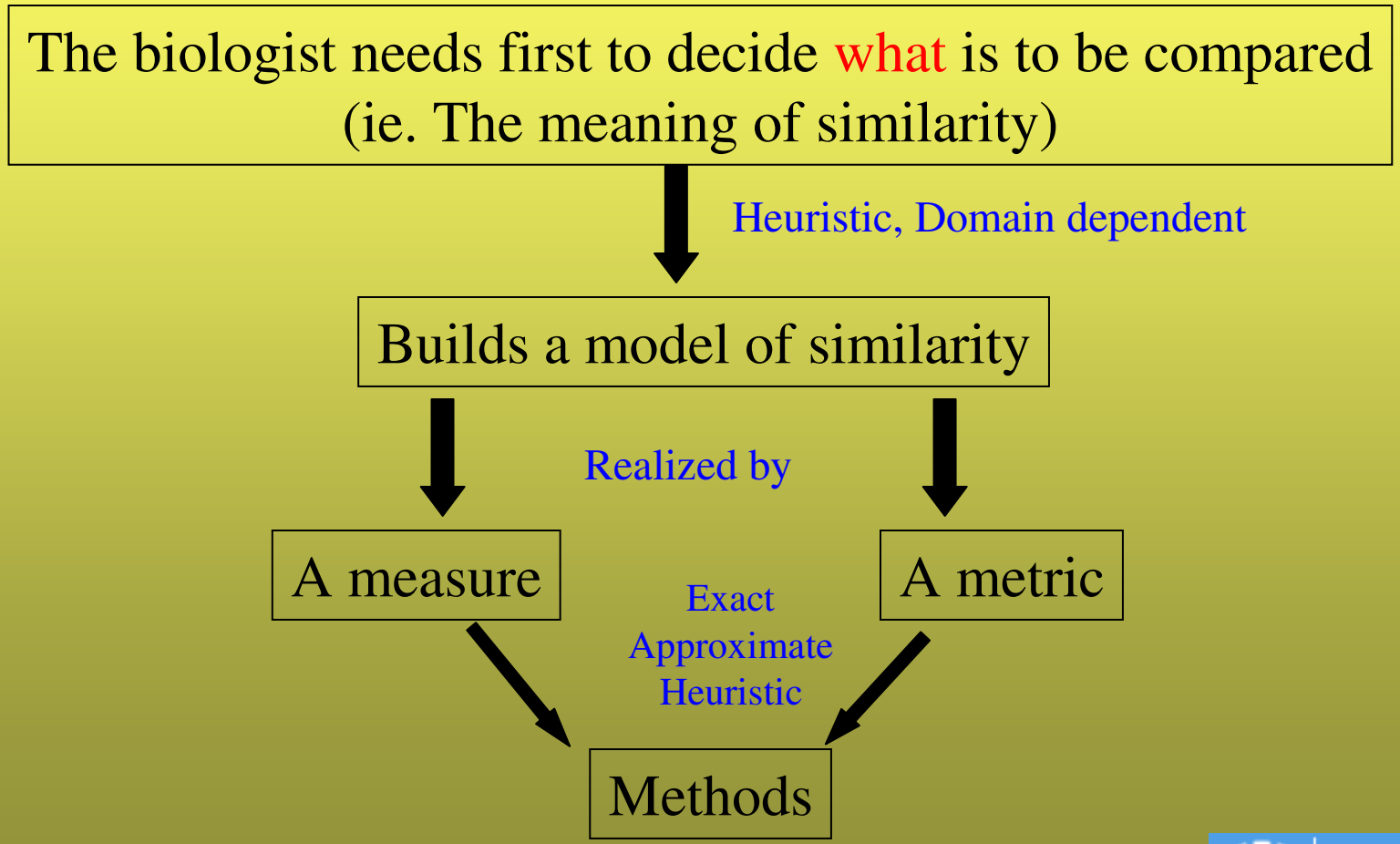
Comparison of Protein Structures: Models, Measures, Metrics and Methods

By Natalio Krasnogor for MIPNETS 20/04/2004

asap automated scheduling optimisation & planning research

The University of Nottingham

What are we comparing?
Models, Measures, Metrics & Methods



Existing Approaches

A variety of structure comparison programs/servers exist:

- SSAP (Orengo & Taylor, 96)
- ProSup (Feng & Sippl, 96)
- DALI (Holm & Sander, 93)
- CE (Shindyalov & Bourne, 98)
- LGA (Zemla, 2003)
- SCOP (Murzin, Brenner, Hubbard & Chothia, 95)
- CATH (Orengo, Mithie, Jones, Jones, Swindells & Thornton, 97}

These are based on:

- Dynamic programming (Taylor, 99)
- Comparison of distance matrices (Holms & Sander, 93,96}
- Maximal common sub-graph detection (Artimiuk, Poirrette, Rice & Willet, 95)
- Geometrical matching (Wu, Schmidler, Hastie & Brutlag, 98)
- Root-mean-square-distances (Maiorov & Crippen, 94 – Cohen & Sternberg,80)
- Other methods (eg. Lackner, Koppensteimer, Domingues & Sippl, 99 – Zemla, Vendruscolo, Moult & Fidelis, 2001)

An excellent survey of various (37 in total) similarity measures can be found in (May, 99)

Note that:

- No consensus on which of these is the best method
- Various difficulties are associated with each.
- They assume that a suitable scoring function can be defined for which optimum values correspond to the best possible structural match between two structures
- RMSD based, eg., may have numerical instabilities problems
- Some methods cannot produce a proper ranking due to:

 - ambiguous definitions of the similarity measures

 or

 -neglect of alternative solutions with equivalent similarity values.

An often over-looked problem associated with some of the established comparison methods:

Whilst similarity can at least (but not only) be measured by the minimum RMSD between two structures and also by their number of equivalent residues these two measures are not completely (in)dependent , i.e. the optimization of one does not necessarily follow from the optimization of the other.

For example:

• ProSup (Feng & Sippl, 96) optimizes the number of equivalent residues with the RMSD being an additional constraint (and not another search dimension).

• DALI (Holm & Sander, 93) combines various derived measures into one value, effectively transforming a multi-objective problem into a (weighted) single objective one.

The structural comparison problem should be, ideally, treated as a truly multiobjective.

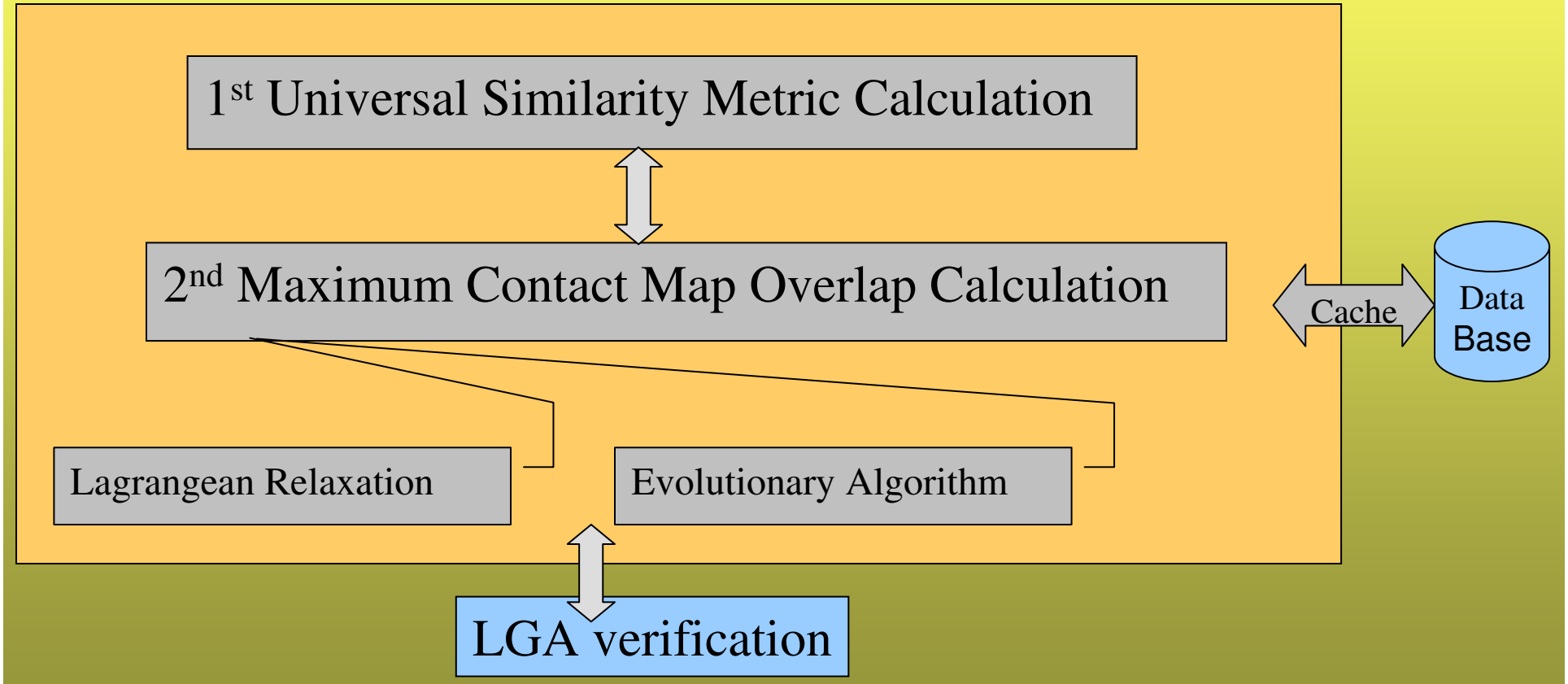
Thus, three main approaches for structural comparison:

• One of the protein structures is fixed and the second is rotated and translated as a rigid body to minimize its RMSD from the first structure (Kabsch, 79).

• A similarity measure based on distance matrices (Holms & Sander, 93) -related to the one we present here but not entirely identical-

• A similarity based on contact map overlaps is the only one of the three approaches that does not require a pre-calculated set of residues equivalences as one of the goals of the method is in fact to determine the best equivalences (Godzick, Skolnick & Kolinski, 1992)

A New Protocol for Protein Structure Comparison



Measuring the Similarity of Protein Structures by Means of the Universal Similarity Metric

(Krasnogor & Pelta, 2004 in *Bioinformatics*)

No need to decide *a priory* which biological model to assume! (the *what* question)

USM approximates every possible similarity metric

USM introduced in (Li, Badger, Chen, Kwon, Kearney & Zhang, 2001)

USM refined in (Li, Chen, Li, Ma & Vitanyi, 2003)

At the core of USM lies the concept of Kolmogorov Complexity.
The Kolmogorov complexity K(.) of an object o is defined by the length of the shortest program for a Universal Turing Machine U that is needed to output o.
That is:

$$K(o) = \min \{ |P|, P \text{ is a program and } U(P)=o\} \quad (1)$$

A related measure is the conditional Kolmogorov complexity of o_1 given o_2:

$$K(o\_1|o\_2) = \min \{ |P|, P \text{ is a program and } U(P,o\_2)=o\_1 \} \quad (2)$$

and measures how much information is needed to produce object 1 if we know object 2.

It is possible to show that the Information Distance between two objects is equivalent (up to a logarithmic additive term) to:

$$ID(o\_1,o\_2) = \max \{ K(o\_1|o\_2), K(o\_2|o\_1) \} \quad (3)$$

The Universal Similarity Measure, as introduced in (Lin, Chen, Lin, Ma & Vitanyi, 2003) is a proper metric, it is universal and also normalized.

The metric is formally defined as:

$$d(o\_1,o\_2) = \frac{\max \{ K(o\_1|o\_2^*), K(o\_2|o\_1^*) \}}{\max\{ K(o\_1),K(o\_2) \}} \qquad (4)$$

where o_1^* ,o_2^* indicates a shortest program for o_1 , o_2 respectively.

Using Eq. (4) we can produce a matrix with the USM distance between proteins o_1 and o_2 for all o_1,o_2 in a set to be compared.

How do we actually compute d(.,.)?

- The universality of the USM is paid by non-computability, that is, Kolmogorov complexity is non-computable but only upper-semi computable.

- We need to approximate d(.,.) by approximating K(.):

- Each protein is encoded as a string s and K(s) is approximated by the size (i.e. number of bytes) of the compressed string zip(s), that is, $K(s) \sim |zip(s)|$ (5)

- In (Li & Vitanyi, 97) it is shown that algorithmic information is symmetric, hence we can also approximate $K(o_1|o_2)$ by $K(o_1 + o_2)-K(o_2)$ where + denotes string concatenation and K(.) is estimated as mentioned above.

From the PDB we can obtain detailed structural information

Chain 1AI9:A							
bond	total #	average	stddev	min	at	max	at
C-N	180	1.32	0.019	1.27	VAL 6	1.38	ASN 123
C-N (PRO)	11	1.33	0.019	1.29	PRO 68	1.36	PRO 160
C-O	192	1.25	0.022	1.19	ASN 124	1.33	GLN 165
CA-C	184	1.52	0.022	1.47	LEU 121	1.58	ILE 8
CA-C (GLY)	8	1.54	0.016	1.52	GLY 20	1.57	GLY 55
CA-CB	133	1.53	0.032	1.4	GLU 174	1.62	ASP 105
CA-CB (ALA)	7	1.53	0.019	1.5	ALA 93	1.56	ALA 16
CA-CB (I,T,V)	44	1.56	0.026	1.5	VAL 6	1.61	THR 147
N-CA	173	1.47	0.023	1.42	ASP 71	1.54	TRP 189
N-CA (GLY)	8	1.47	0.013	1.45	GLY 20	1.49	GLY 180
N-CA (PRO)	11	1.47	0.02	1.44	PRO 15	1.5	PRO 152

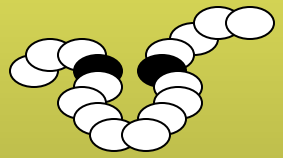
Source: http://www.rcsb.org/pdb

But PDB also contains other information which is not relevant to structural related activities (e.g.. the lab name where the X-ray Crystallography, NMR was done).

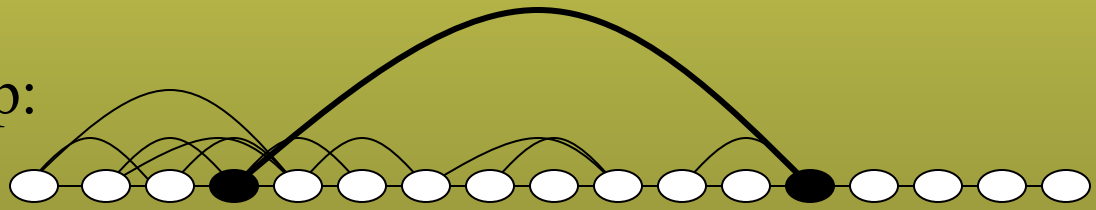
So, instead of using the whole PDB file of a protein in order to compute its USM we only use a contact map:

A protein:

Its structure:



The structure's contact map:



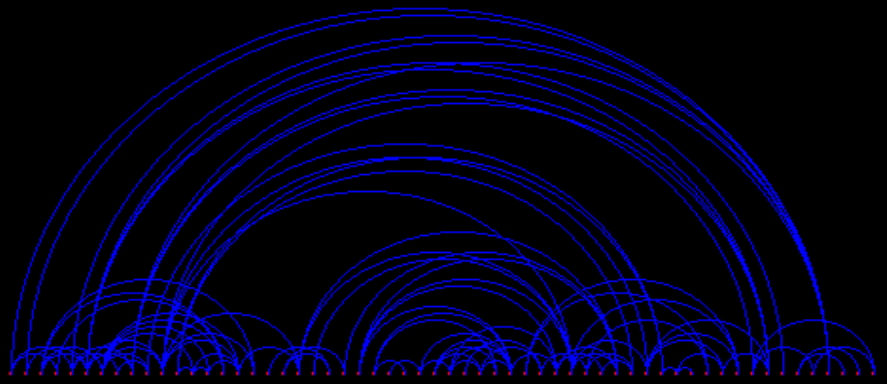
Formally:

A CM is a concise representation of a protein's native three-dimensional structure. A CM is specified by a 0-1 matrix S, with entries indexed by pairs of protein residues

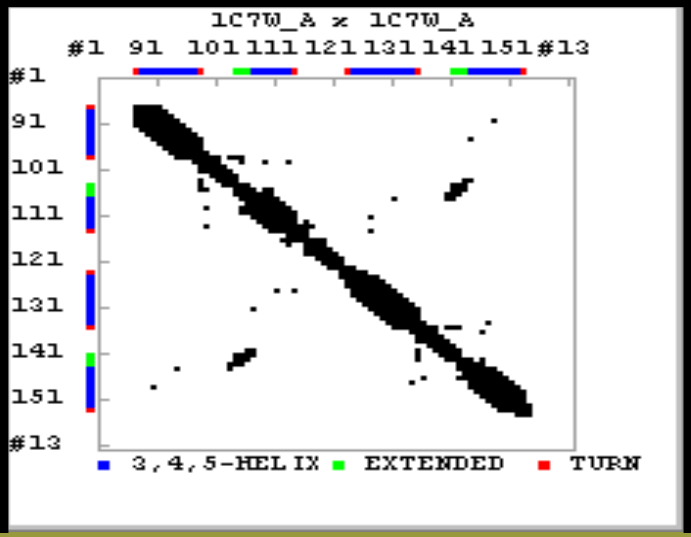
$$S_{i,j} = \begin{cases} 1 \textit{ if residue i and j are in contact} \\ \\ 0 \textit{ otherwise} \end{cases}$$

Residues i and j are said to be *in contact* if they lie within *R* Angstroms from each other in the protein's native fold.

R is called the *threshold* of the contact map



1C7W.PDB

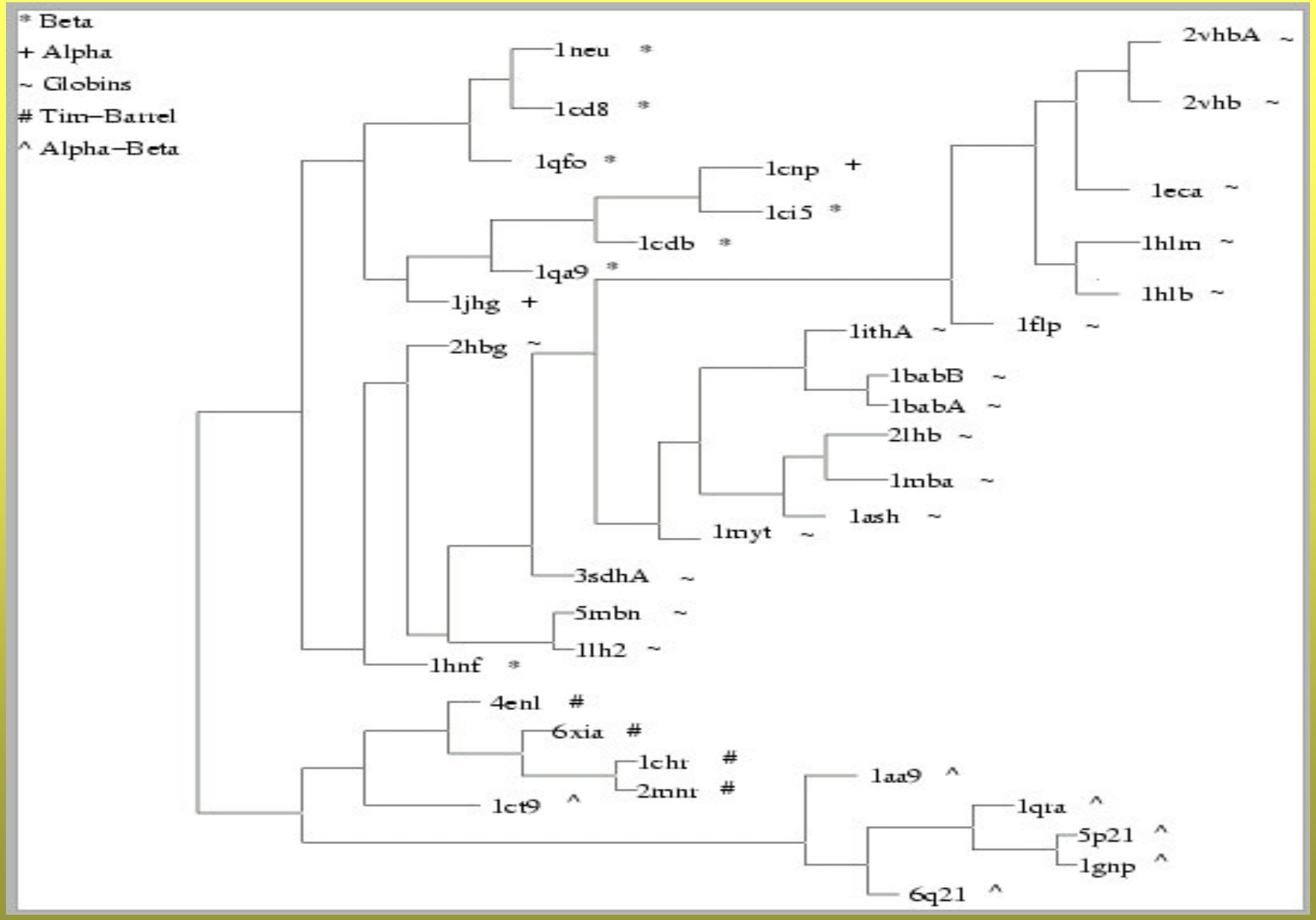

1C7W_A x 1C7W_A
#1 91 101 111 121 131 141 151 #13
#1
91
101
111
121
131
141
151
#13
3,4,5-HELIX
EXTENDED
TURN

Example with the Chew-Kedem data set

• This data set was used in (Chew & Kedem, 2002) to assess the quality of a newly proposed method to measure consensus shapes.

• These are 36 medium size proteins of 5 different families
  - globins {1eca, 5mbn, 1hlb, 1hlm, 1babA, 1babB, 1ithA, 1mba, 2hbg, 2lhb, 3sdhA, 1ash, 1flp, 1myt, 1lh2, 2vhbA, 2vhb}
  - alpha-beta {1aa9, 1gnp, 6q21, 1ct9, 1qra, 5p21}
  - tim-barrels {6xia, 2mnr, 1chr, 4enl}
  - all beta {1cd8, 1ci5, 1qa9, 1cdb, 1neu, 1qfo, 1hnf}
  - and alpha {1cnp,1jhg}

• Protein 2vhb was repeated two times (as 2vhb and 2vhbA) to check whether the USM detects that the two are identical and induces a cluster where both appear together.



asap automated scheduling optimisation & planning research

Comparison of Protein Structures: Models, Measures, Metrics and Methods

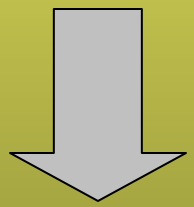
By Natalio Krasnogor for MIPNETS 20/04/2004

The University of Nottingham

So, USM allows us to measure the similarity of protein structures without answering the "what?" question

But…

it does not tell us how these structures are (di)similar



We use Maximum Contact Map Overlap for that!

A Comparison of Computational Methods for the Maximum Contact Map Overlap of Protein Pairs

(Krasnogor, Lancia, Zemla, Hart, Carr, Hirst & Burke, to be submitted)

- Protein similarity can be computed by *aligning* the two contact maps of a pair of proteins

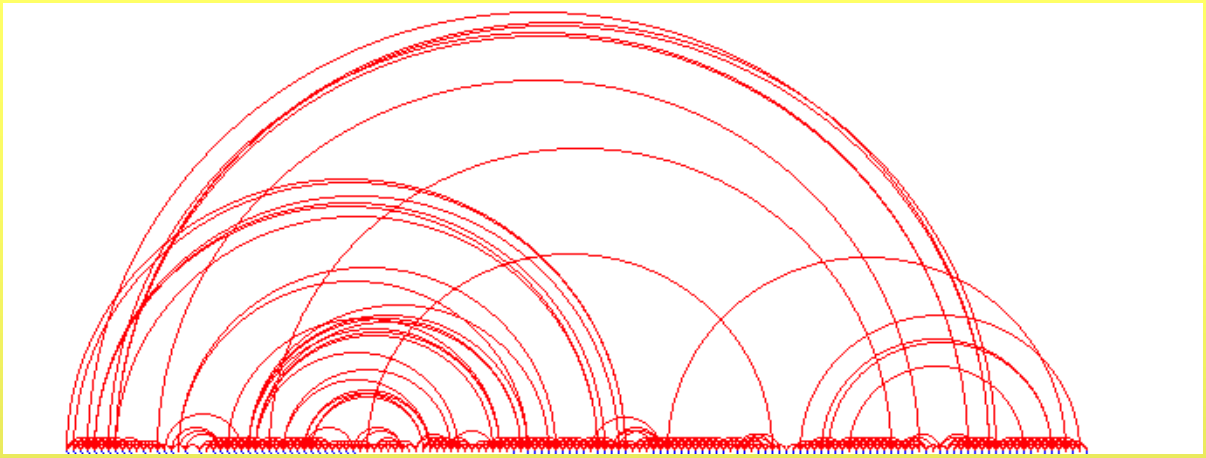
- An *alignment* of two proteins is a pairing of amino acids between them

1ash

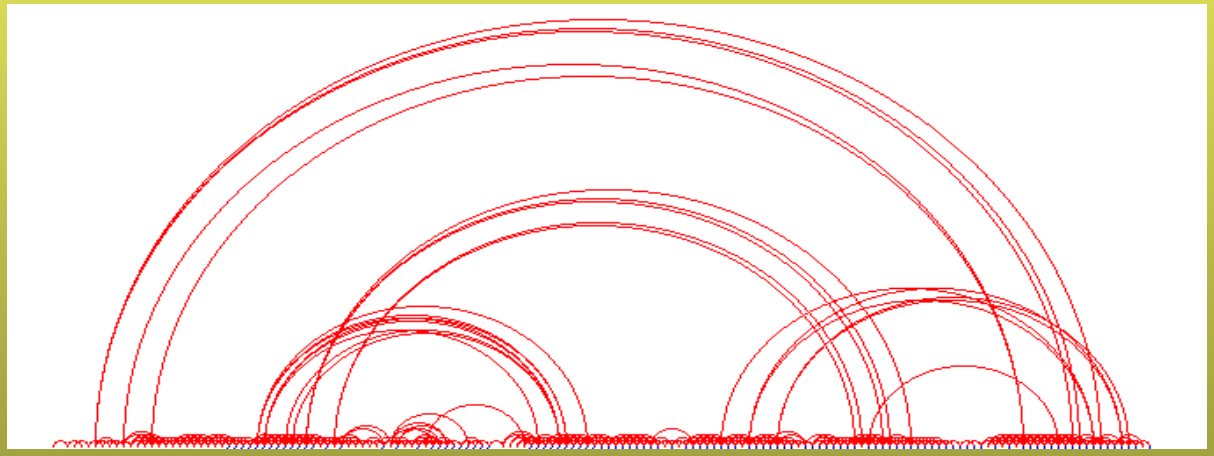
1hlm

Two related proteins taken from the PDB which share a 6 helices structural motif.

Contact maps of as a graph in which each contact between two residues corresponds to an edge

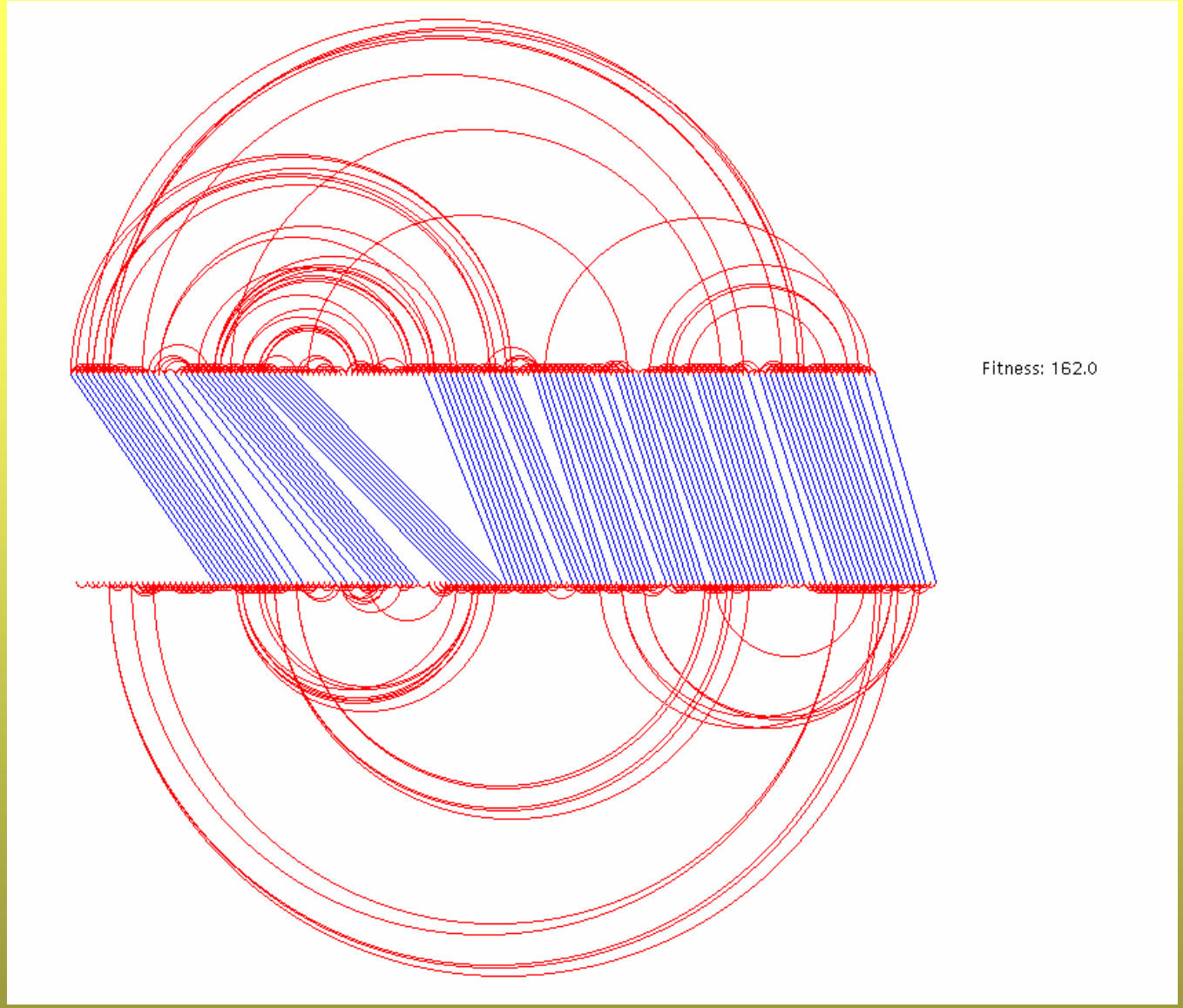


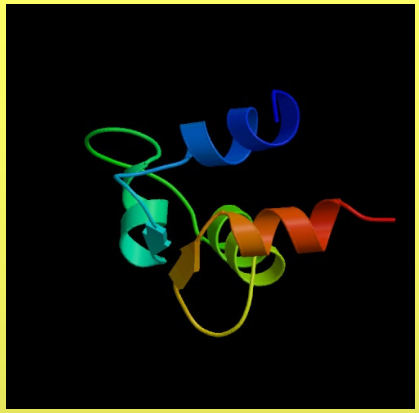
1ash contact map



1hlm contact map

A candidate alignment between the contact maps of these protein structures.

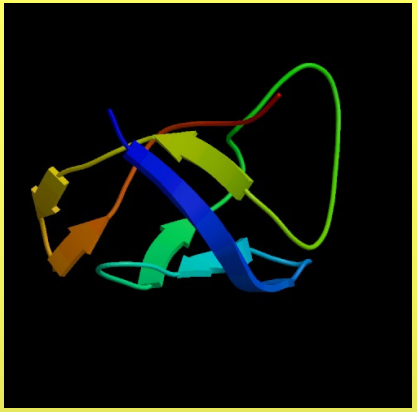




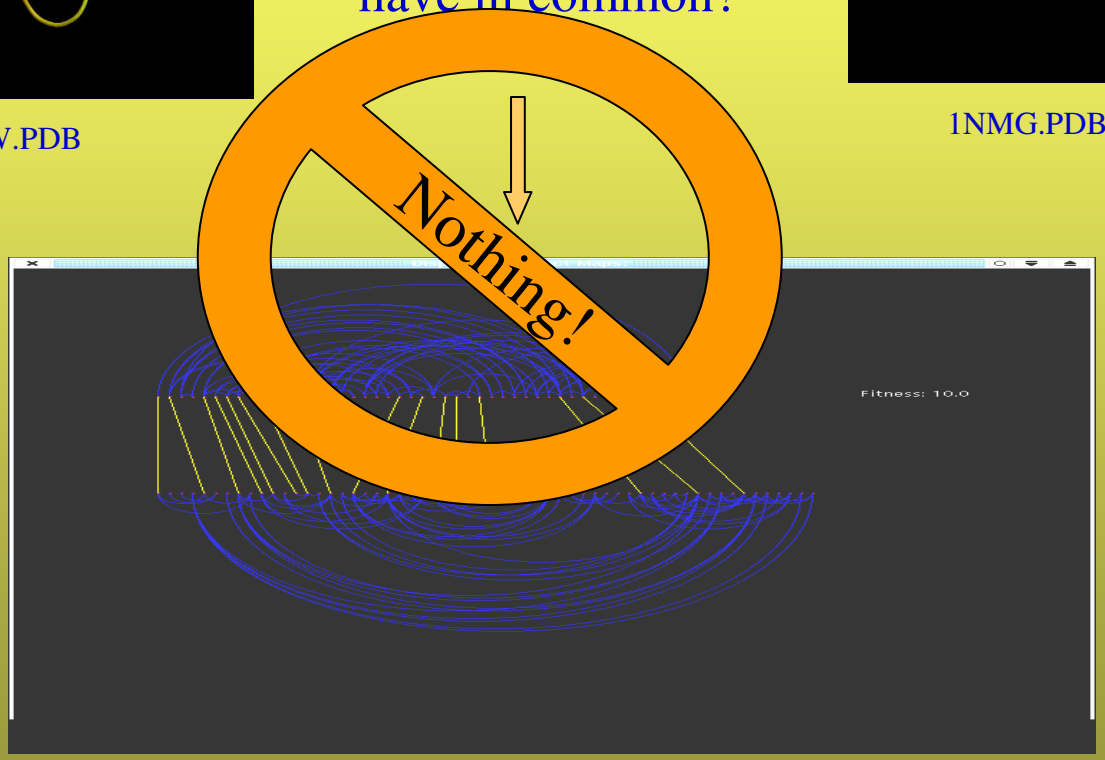
1C7W.PDB

What do these structures

have in common?



1NMG.PDB



One possible alignment between the contact maps of 1C7W and 1NMG

The Maximum Contact Map Overlap Problem can be modelled with the following IP formulation (Caprara & Lancia, 2002):

$$\max \quad \sum_{e \in E_1} \sum_{f \in E_2} y_{ef}$$

s.t.

$$\sum_{j \in V_1^+(i)} y_{(i,j)(u,v)} \leq x_{iu} \qquad \forall i \in V_1, \forall (u,v) \in E_2$$

$$\sum_{j \in V_1^-(i)} y_{(j,i)(u,v)} \leq x_{iv} \qquad \forall i \in V_1, \forall (u,v) \in E_2$$

$$\sum_{v \in V_2^+(u)} y_{(i,j)(u,v)} \leq x_{iu} \qquad \forall u \in V_2, \forall (i,j) \in E_1$$

$$\sum_{v \in V_2^-(u)} y_{(i,j)(v,u)} \leq x_{iv} \qquad \forall u \in V_2, \forall (i,j) \in E_1$$

$$\sum_{[u,v] \in F} x_{uv} \leq 1 \qquad \forall F \in \mathcal{I}$$

$$x, y \in \{0,1\}.$$

• This problem formulation is suitable for a robust and fast Lagrangean relaxation (LR) method.

•The MAX-CMO has also been tackled with a Memetic Algorithm (MA), which is a hybrid evolutionary-local search algorithm.

• LR delivers the best known solutions to these alignments, in most cases the optimal ones. For those that are not optimal we can compute the gap between the optimal and the best result.

•MA delivers sub-optimal solutions but lots of them, this allows the end-user to pick the one that is more biologically meaningful and relevant

• MAX-CMO is the only model for which exact optimal solutions and certifiably sub-optimal solutions can be obtained.

•We validated our two-tier protocol with Local-Global alignment (LGA) (Zemla, 2003)

•LGA has been itself validated in several CASP competitions as the method to assess the similarity between the model structures and their targets

• LGA is an accepted method of similarity

•The scoring function based on two measures:

- LCS, stands for the Longest Continuous Segment
- GDT, stands for Global Distance Test

• LCS is designed to capture the local similarities between two structures by finding the longest subset of contiguous residues that can be rigidly superimposed within a pre-fixed RMSD threshold.

• The reference atoms between residues are the C_α atoms.

• Considers all the possible contiguous sub-segments of residues until it finds the one which deviates minimally from the RMSD considered.

• The LCS measure can be efficiently computed with a dynamic programming (Kabsch, 79).

•This is an exact but local evaluation of structural similarity.

• GDT tries to obtain the largest set of equivalent residues that fit within a fixed distance cutoff and that are not necessarily contiguous.

• This is a combinatorial problem in nature and as such can only be solved approximately.

• GDT evaluates a selected but large number of superpositions

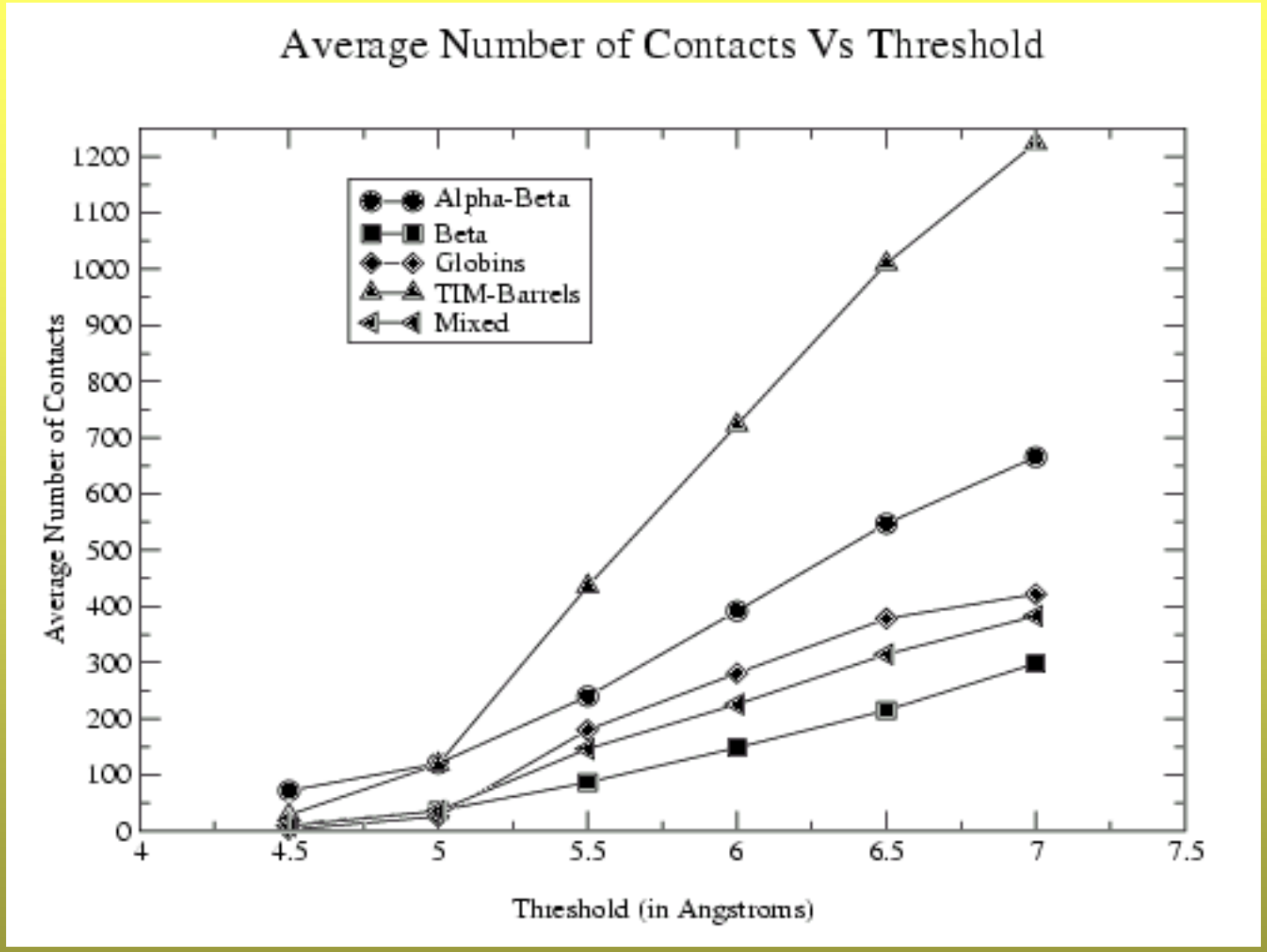
• GDT provides global information about the similarity regions of the two proteins.

•LCS algorithm identify local regions of similarity between proteins,

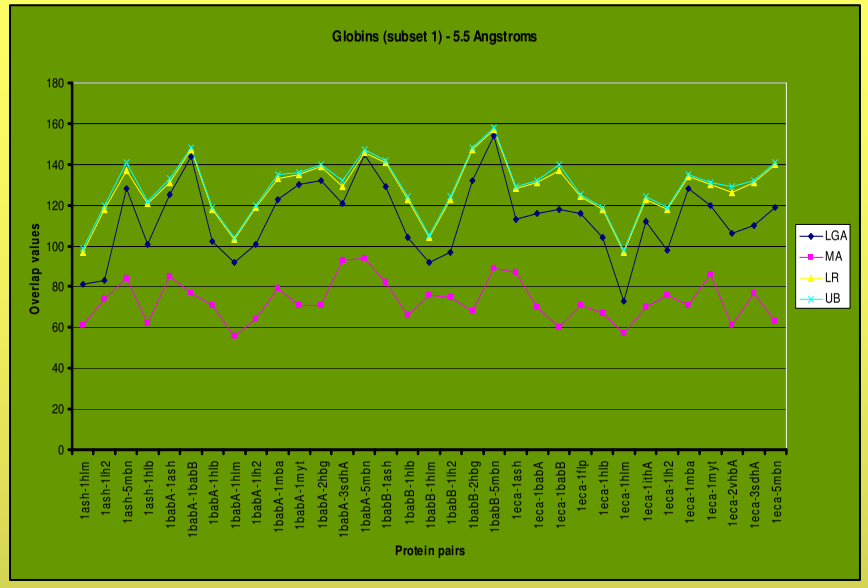
•GDT arise information from anywhere in the structure.

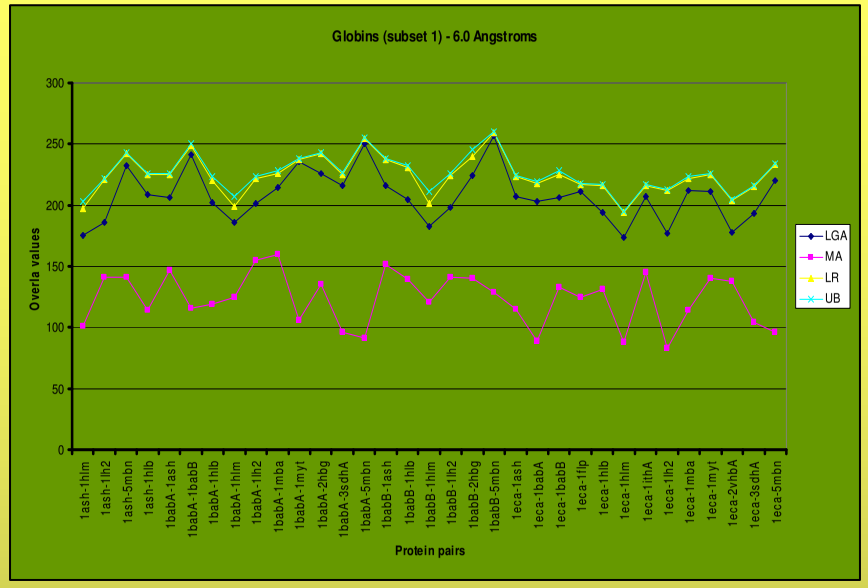
Results

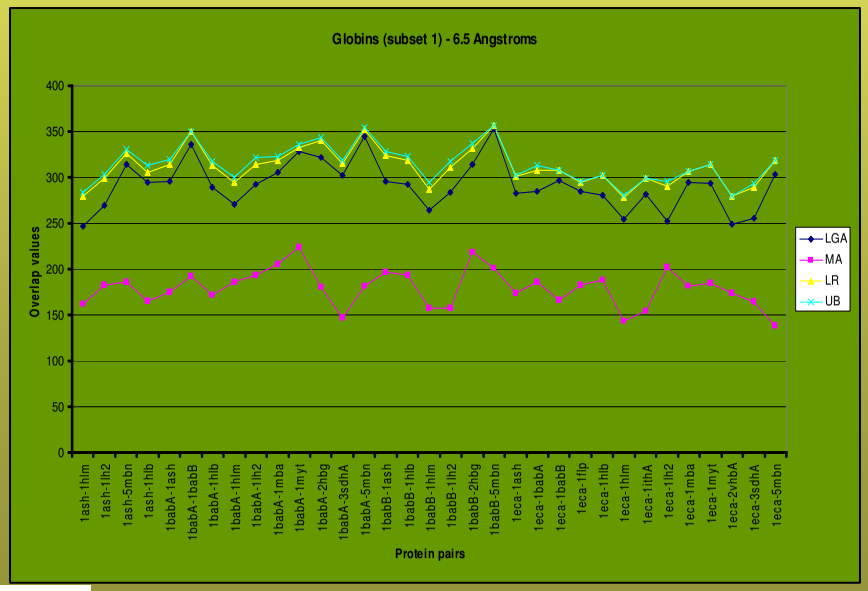
Protein Family	Avg. Number of Residues
Alpha-Beta	223.8
Beta	108.7
Globins	147.1
TIM-barrels	388.5
Mixed	136.4

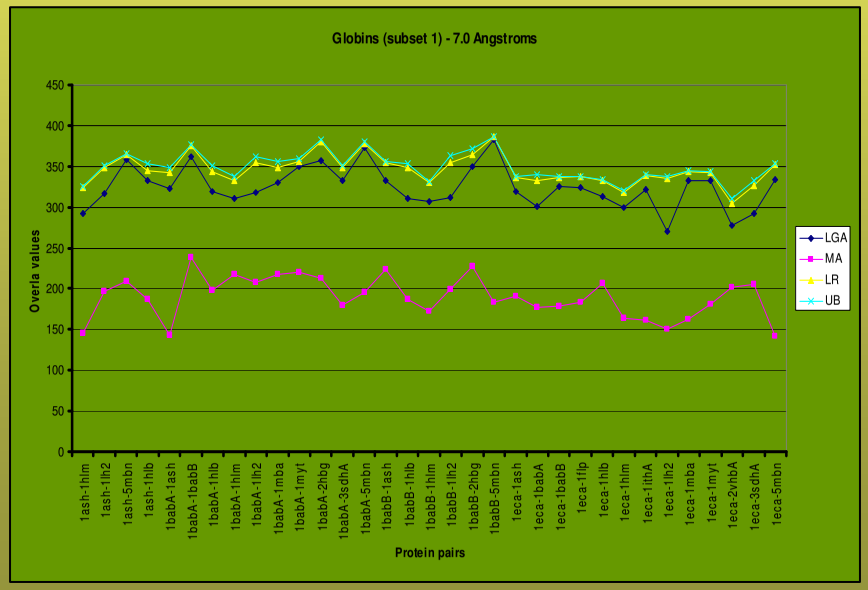

Average Number of Contacts Vs Threshold
Alpha-Beta
Beta
Globins
TIM-Barrels
Mixed
Average Number of Contacts
Threshold (in Angstroms)
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
4
4.5
5
5.5
6
6.5
7
7.5

Globins (subset 1)









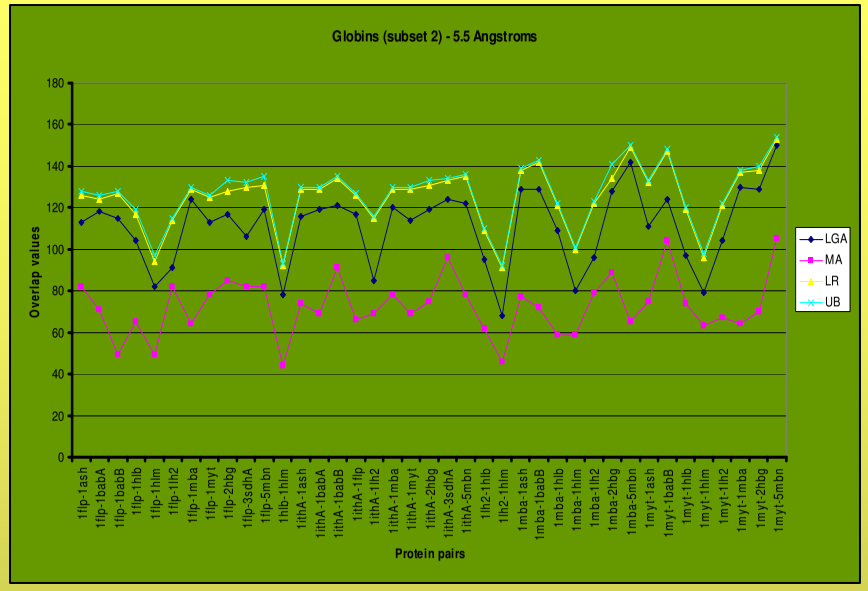
asap research
automated scheduling optimisation & planning

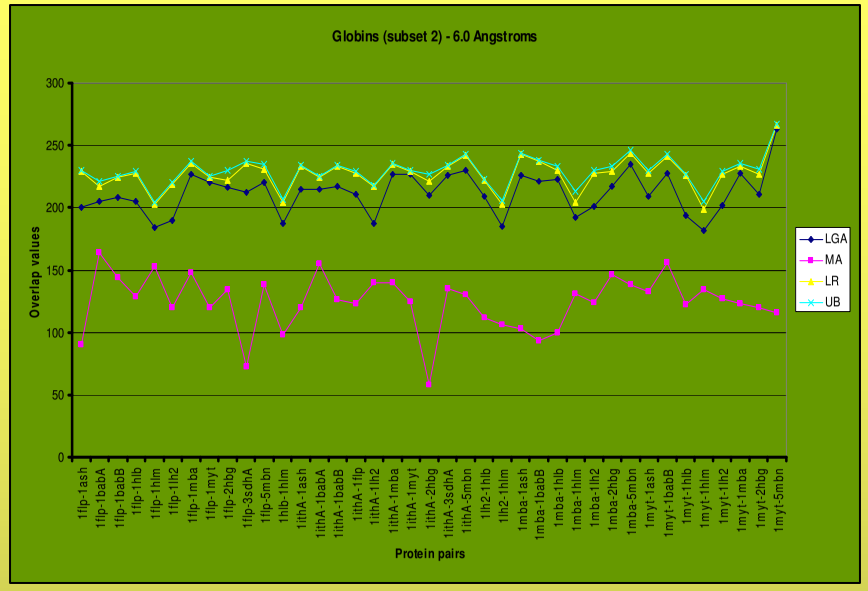
Comparison of Protein Structures: Models, Measures, Metrics and Methods

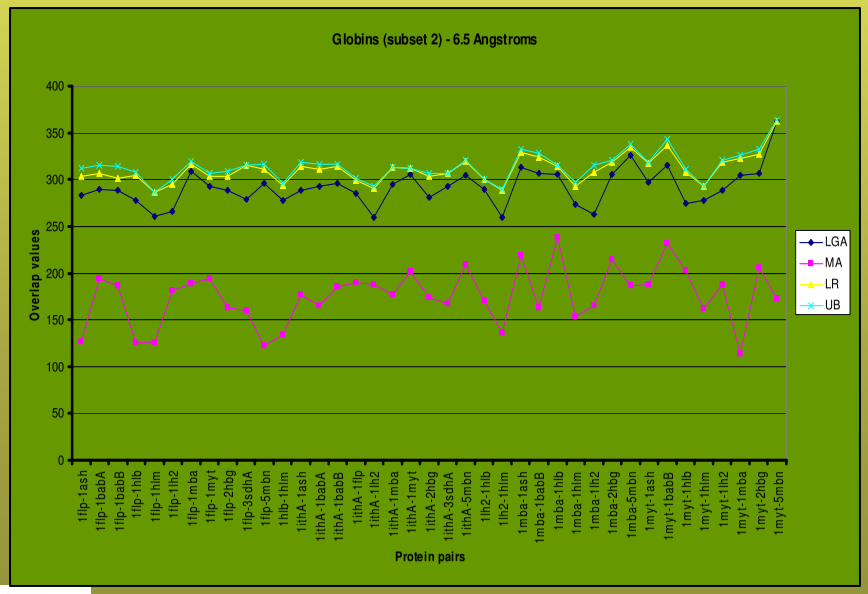
By Natalio Krasnogor for MIPNETS 20/04/2004

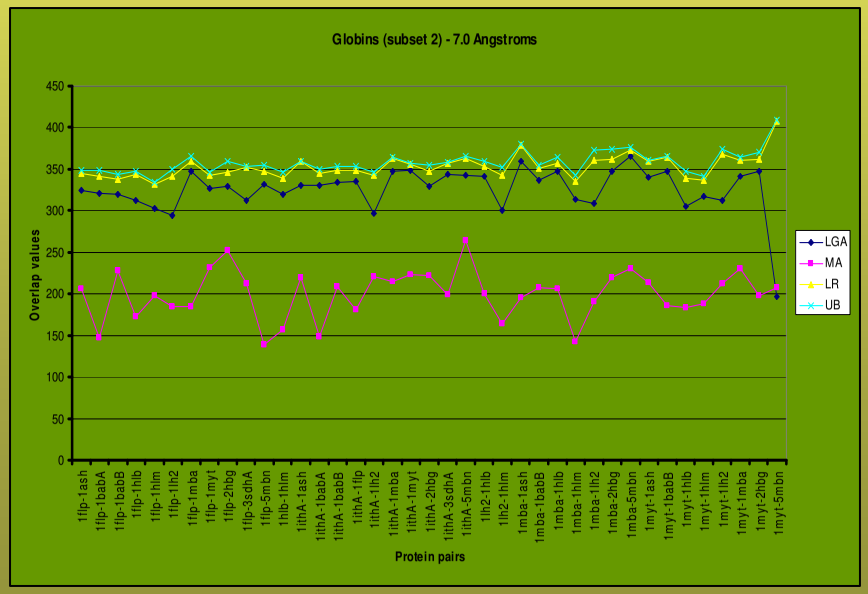
The University of Nottingham

Globins (subset 2)

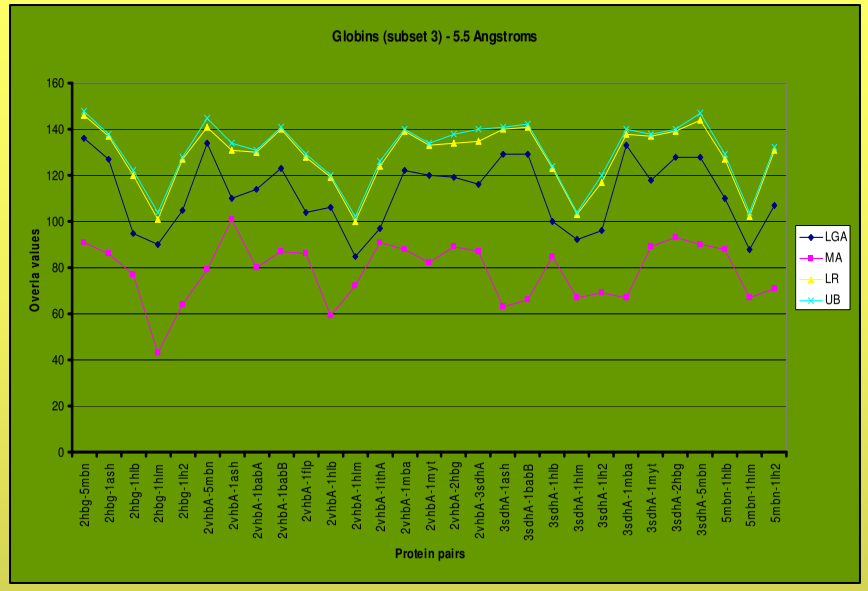


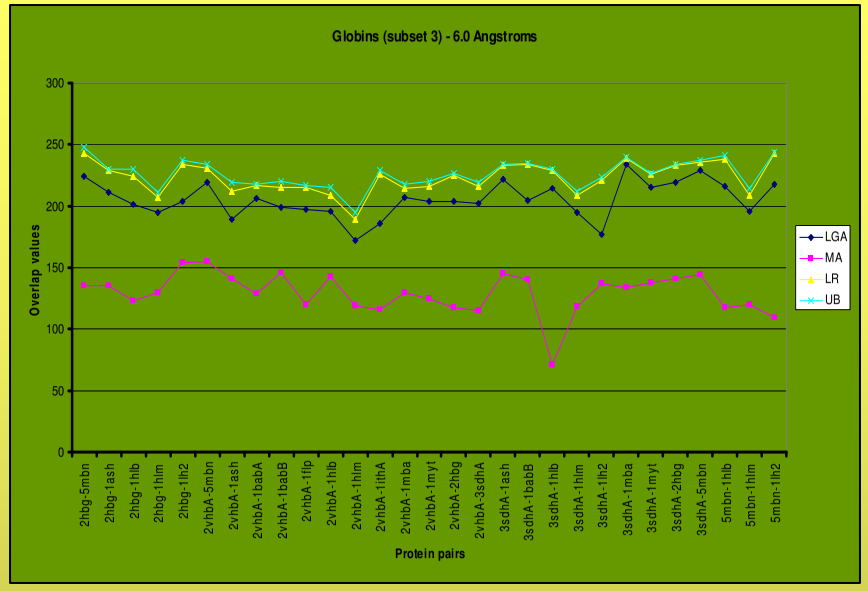


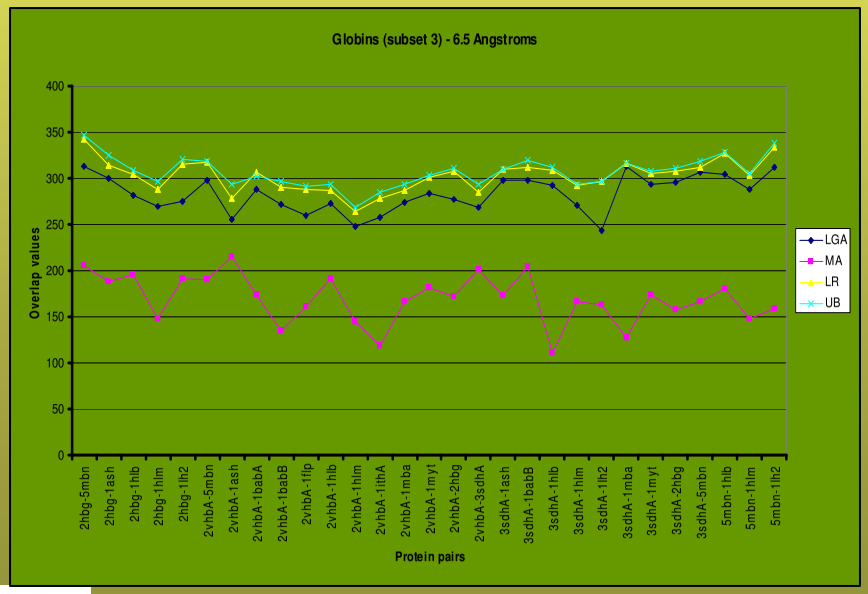


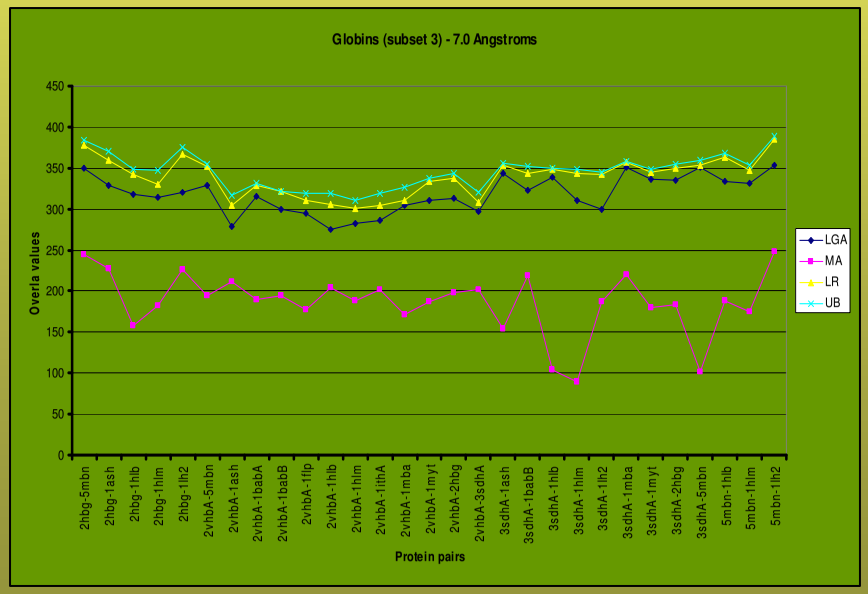


Globins (subset 3)









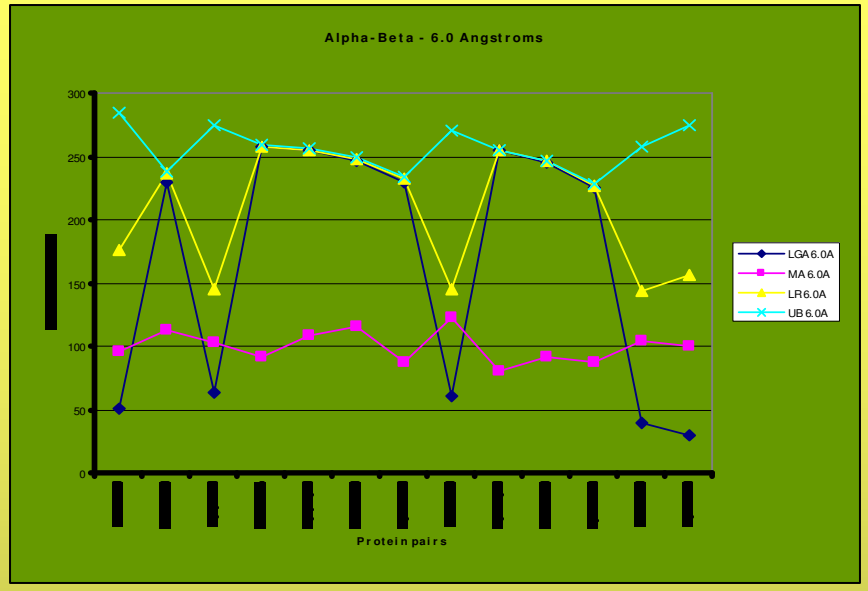
asap automated scheduling optimisation & planning research

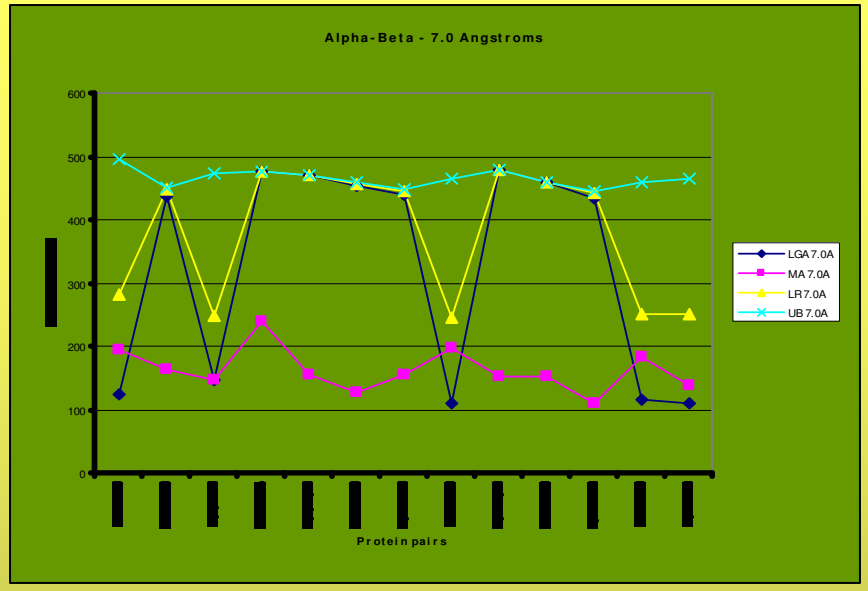
Comparison of Protein Structures: Models, Measures, Metrics and Methods

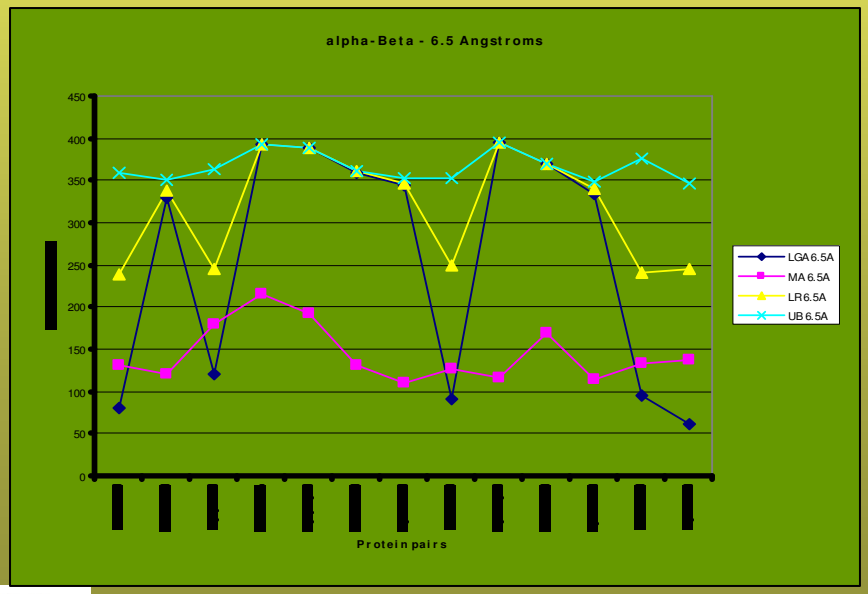
By Natalio Krasnogor for MIPNETS 20/04/2004

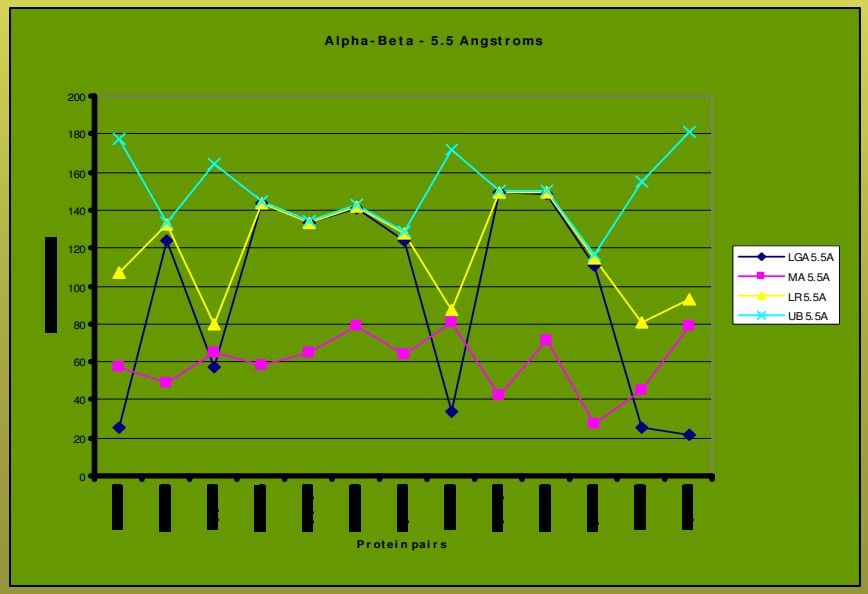
The University of Nottingham

Alpha-Beta









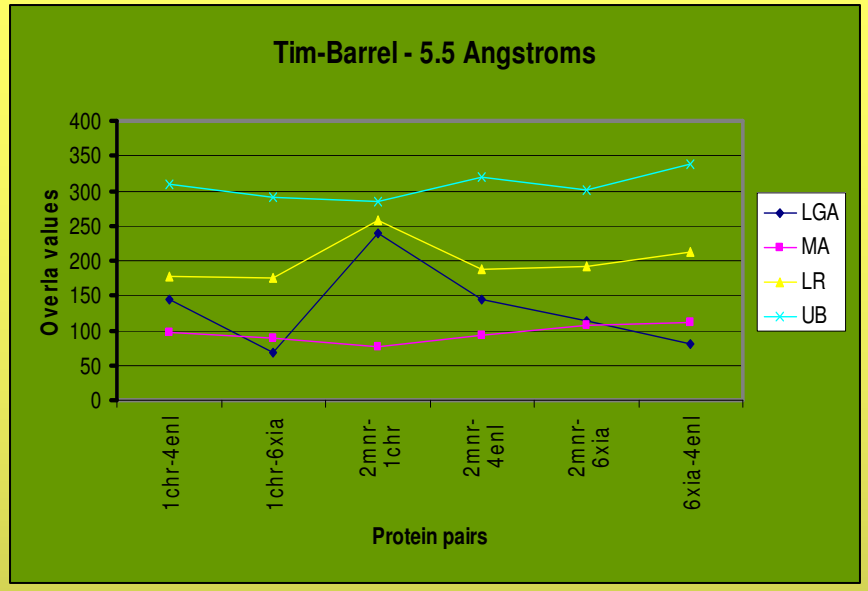
asap automated scheduling optimisation & planning research

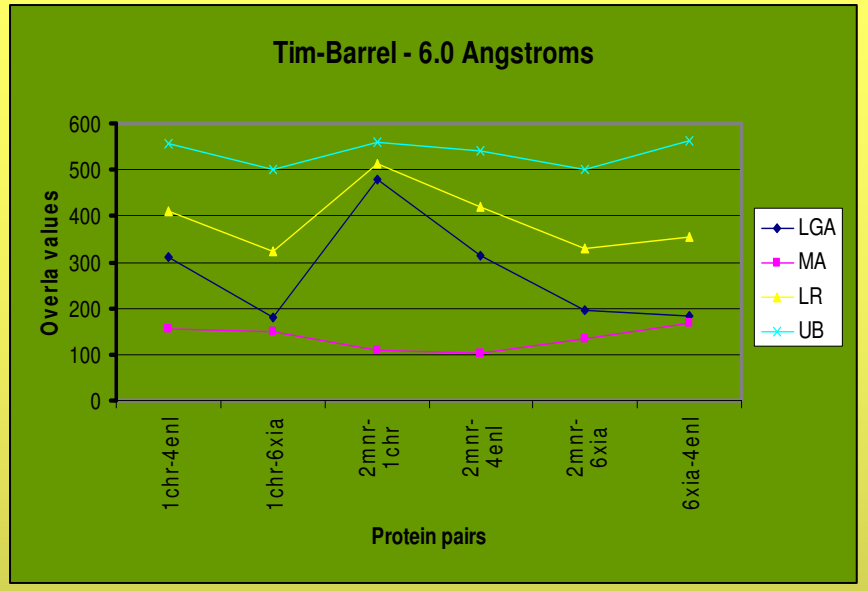
Comparison of Protein Structures: Models, Measures, Metrics and Methods

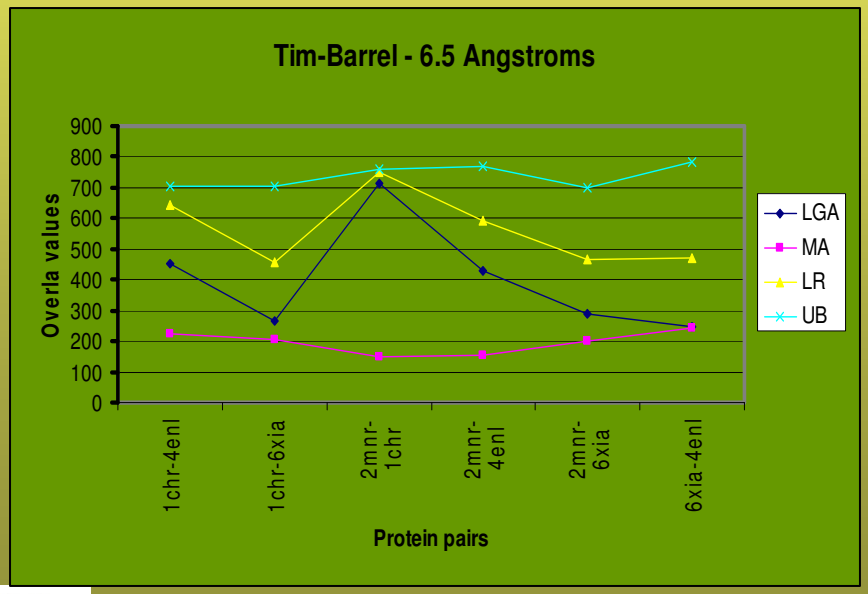
By Natalio Krasnogor for MIPNETS 20/04/2004

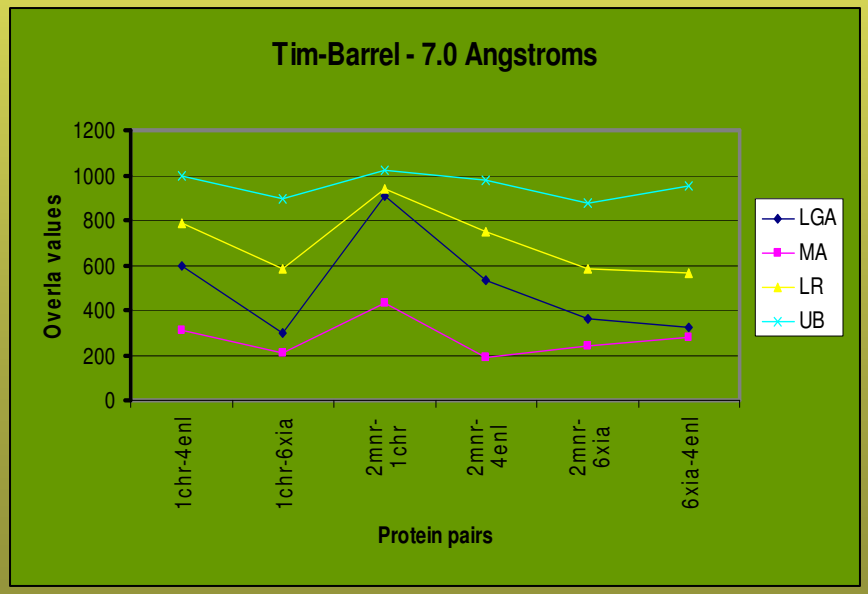
The University of Nottingham

TIM-barrel

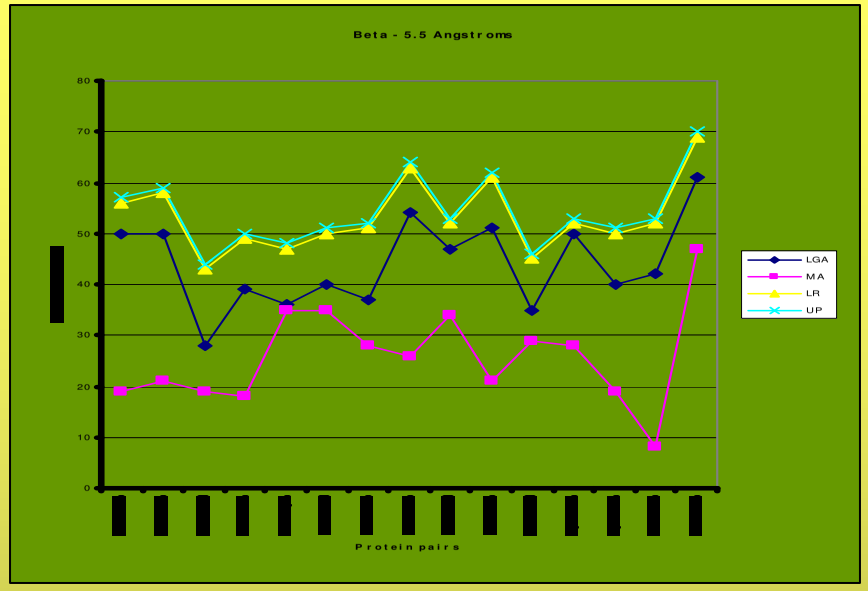

Tim-Barrel - 5.5 Angstroms
Overla values
400
350
300
250
200
150
100
50
0
1chr-4enl
1chr-6xia
2mnr-1chr
2mnr-4enl
2mnr-6xia
6xia-4enl
Protein pairs
LGA
MA
LR
UB

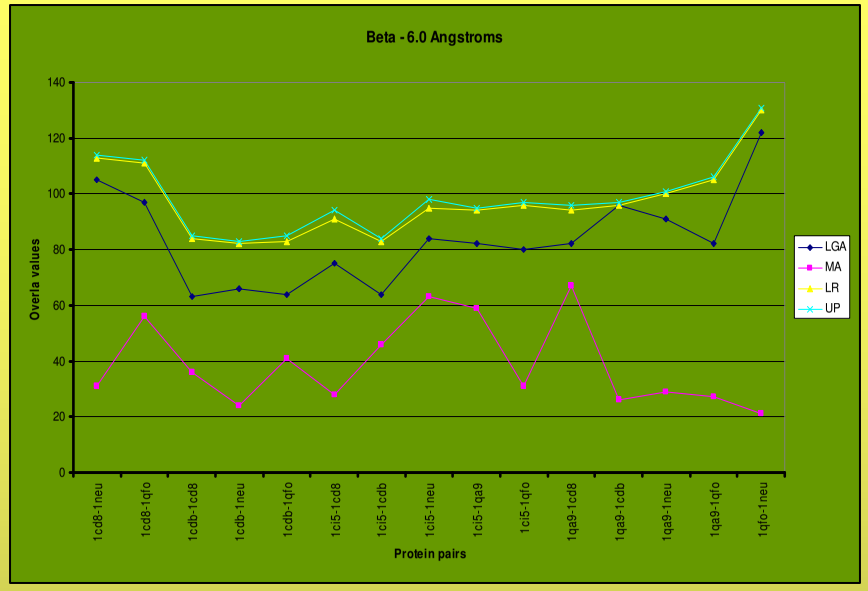

Tim-Barrel - 6.0 Angstroms
Overla values
600
500
400
300
200
100
0
1chr-4enl
1chr-6xia
2mnr-1chr
2mnr-4enl
2mnr-6xia
6xia-4enl
Protein pairs
LGA
MA
LR
UB

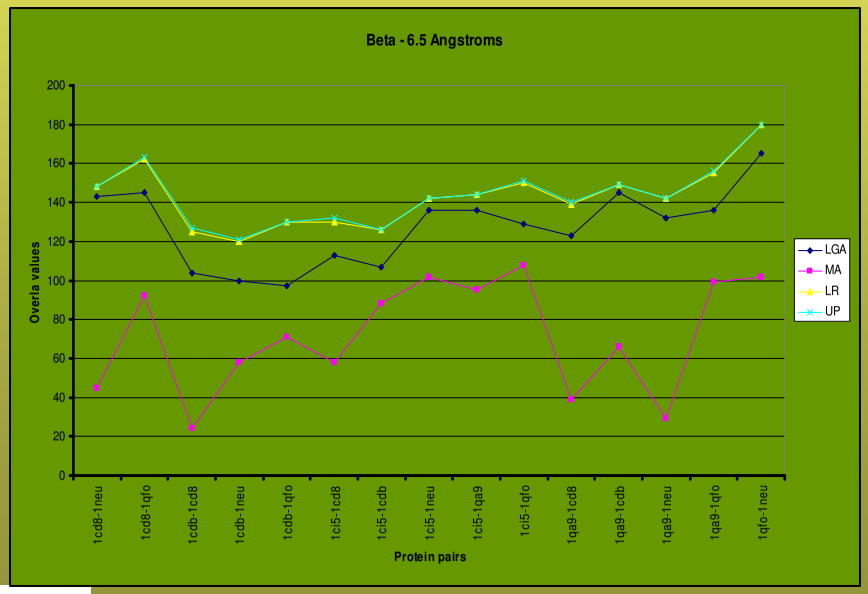

Tim-Barrel - 6.5 Angstroms
Overla values
900
800
700
600
500
400
300
200
100
0
1chr-4enl
1chr-6xia
2mnr-1chr
2mnr-4enl
2mnr-6xia
6xia-4enl
Protein pairs
LGA
MA
LR
UB

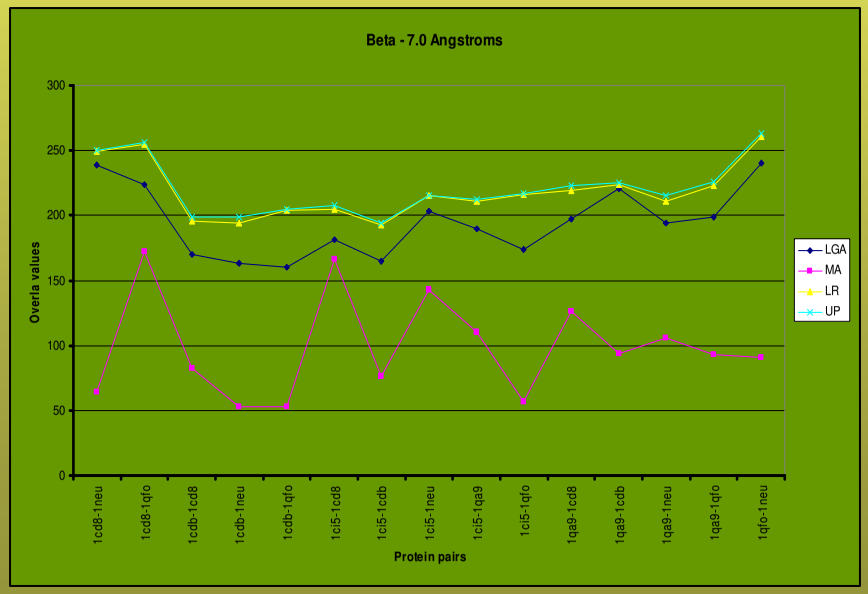

Tim-Barrel - 7.0 Angstroms
Overla values
1200
1000
800
600
400
200
0
1chr-4enl
1chr-6xia
2mnr-1chr
2mnr-4enl
2mnr-6xia
6xia-4enl
Protein pairs
LGA
MA
LR
UB

Beta









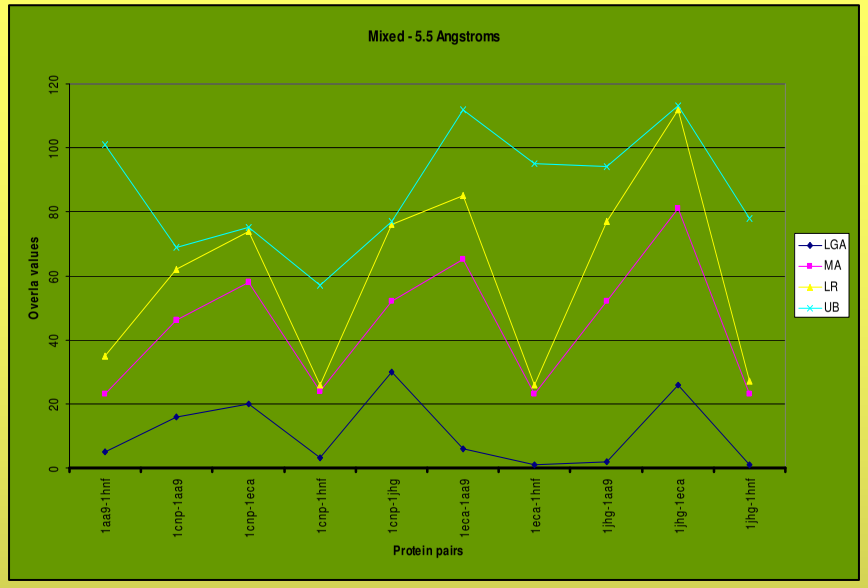
asap research
automated scheduling optimisation & planning

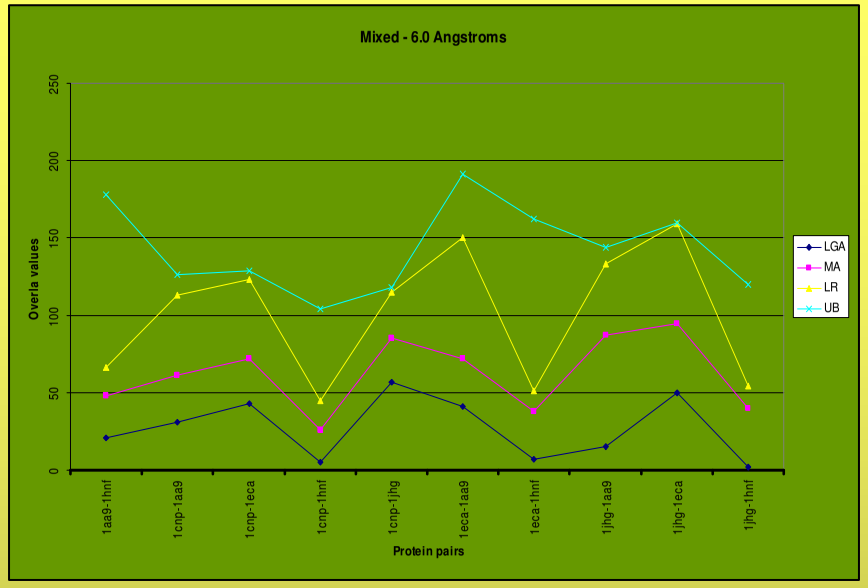
Comparison of Protein Structures: Models, Measures, Metrics and Methods

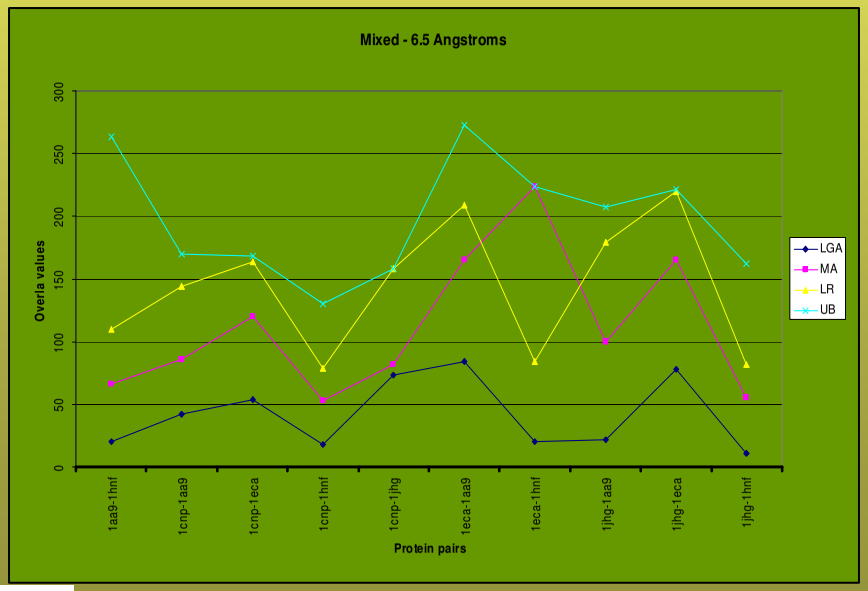
By Natalio Krasnogor for MIPNETS 20/04/2004

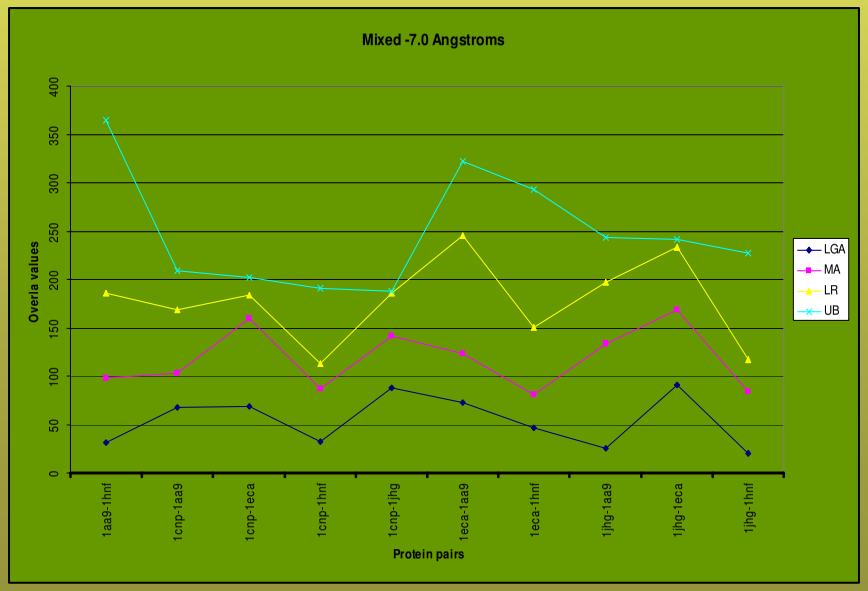
The University of Nottingham

Mixed









asap research
automated scheduling optimisation & planning

Comparison of Protein Structures: Models, Measures, Metrics and Methods

By Natalio Krasnogor for MIPNETS 20/04/2004

The University of Nottingham

	LGA				MA				LR			
Family set	5.5Å	6.0Å	6.5Å	7.0Å	5.5Å	6.0Å	6.5Å	7.0Å	5.5Å	6.0Å	6.5Å	7.0Å
Glob(1) Avg	15.21	18.33	22.72	24.90	55.66	102.75	135.21	159.45	1.39	1.96	3.27	3.42
Std	7.55	8.57	9.71	13.90	12.48	23.64	20.62	24.09	0.78	2.17	2.62	2.68
Glob(2) Avg	15.38	17.41	22.20	32.20	54.07	105.07	138.38	158.58	1.56	2.28	3.30	5.10
Std	6.14	7.64	9.70	32.39	11.45	26.04	28.37	28.19	1.25	2.02	2.78	3.33
Glob(3) Avg	17.93	20.75	23.93	27.68	51.86	96.93	137.86	159.41	1.86	2.96	4.37	6.86
Std	5.72	9.39	10.36	11.85	12.50	19.00	26.56	40.08	1.12	1.89	4.01	4.59
AB Avg	54.76	87.76	107.38	137.69	89.76	156.07	221.53	302.53	31.46	46.38	46.69	84.30
Std	68.87	112.11	134.46	174.99	17.69	18.41	27.52	29.22	40.35	60.10	57.90	109.09
Beta Avg	10.20	15.00	16.00	25.80	28.40	59.53	71.66	121.40	1.00	1.40	0.60	2.06
Std	3.23	6.05	8.03	11.21	10.06	21.98	24.82	38.99	0.0	0.73	0.73	1.48
Tim Avg	175.33	259.83	337.00	450.00	212.00	400.33	540.83	675.66	107.33	145.16	174.00	252.83
Std	71.99	103.29	172.54	185.04	13.75	39.51	59.16	67.54	39.88	56.27	117.48	104.42
Mixed Avg	76.10	116.00	155.40	193.80	42.40	80.80	86.00	130.20	27.10	42.30	54.70	70.00
Std	21.85	32.01	48.89	70.47	20.66	32.78	51.22	74.58	26.67	43.23	55.43	59.09

Average and standard deviation gap for each algorithm on the Globins 1,2 & 3 data sets (Glob), Alpha-Beta (AB), Beta, TIM-barrel (Tim) and Mixed data sets. Contact maps thresholds are 5.5, 6.0, 6.5 and 7.0Å.

PrAKSis Web Server

Web Server

PrAKSis = Protein Alignment and Comparison with Kolmogorov Similarities

www *http://www.cs.nott.ac.uk/praksis*

email ***praksis@cs.nott.ac.uk***

Input Parameters

Structure		Contact Map		Similarity	
Atoms	CA	Threshold	7.5 Angstrom	Method	ContactMaps
Model	1	Exclusion	None	Compressor	gzip
Chain	A	Reduction	None	Equation	Rooij
				Grouping	None

The parameters *model* and *chain* are global settings for all structures. It is not possible to change them for single structures yet.

- Selection of model, chain and atom type of PDB files
- Determine contacts maps by specification of threshold
- Calculate Similarity of pairs of contact maps using real world compressors

Conclusions (1)

• We gave mathematical and experimental evidence that USM can be used to measure the structural (di)similarity between proteins

• USM seems to be able to capture other (more heuristically defined) measures of similarity

• However, USM needs to be complemented with a second tier algorithm that can explicitly say what those similarities are

• We use the alignment of contact map, under a model called The Maximum Contact Map Overlap for that purpose

Conclusions (2)

- We have implemented two distinct algorithms for MAX-CMO:
  - Lagrangean Relaxation
  - Memetic Algorithm

- LR gives the best results known for MAX-CMO and tells how close these results are from the optimum solutions

- The MA provides a family of alternative structural overlaps for the end user to assess in the light of biological (rather than mathematical) relevance

- Our results are at least as good as those produced by LGA which is a well established comparison method.

Future Work(1)

- Investigate how to better approximate USM.

- Extend the LGA web-server to report also contact map overlap values.

- Improve the memetic evolutionary algorithm with problem-specific operators designed for the different families of proteins.

- Investigate how to deal with instances consisting of substantially different proteins.

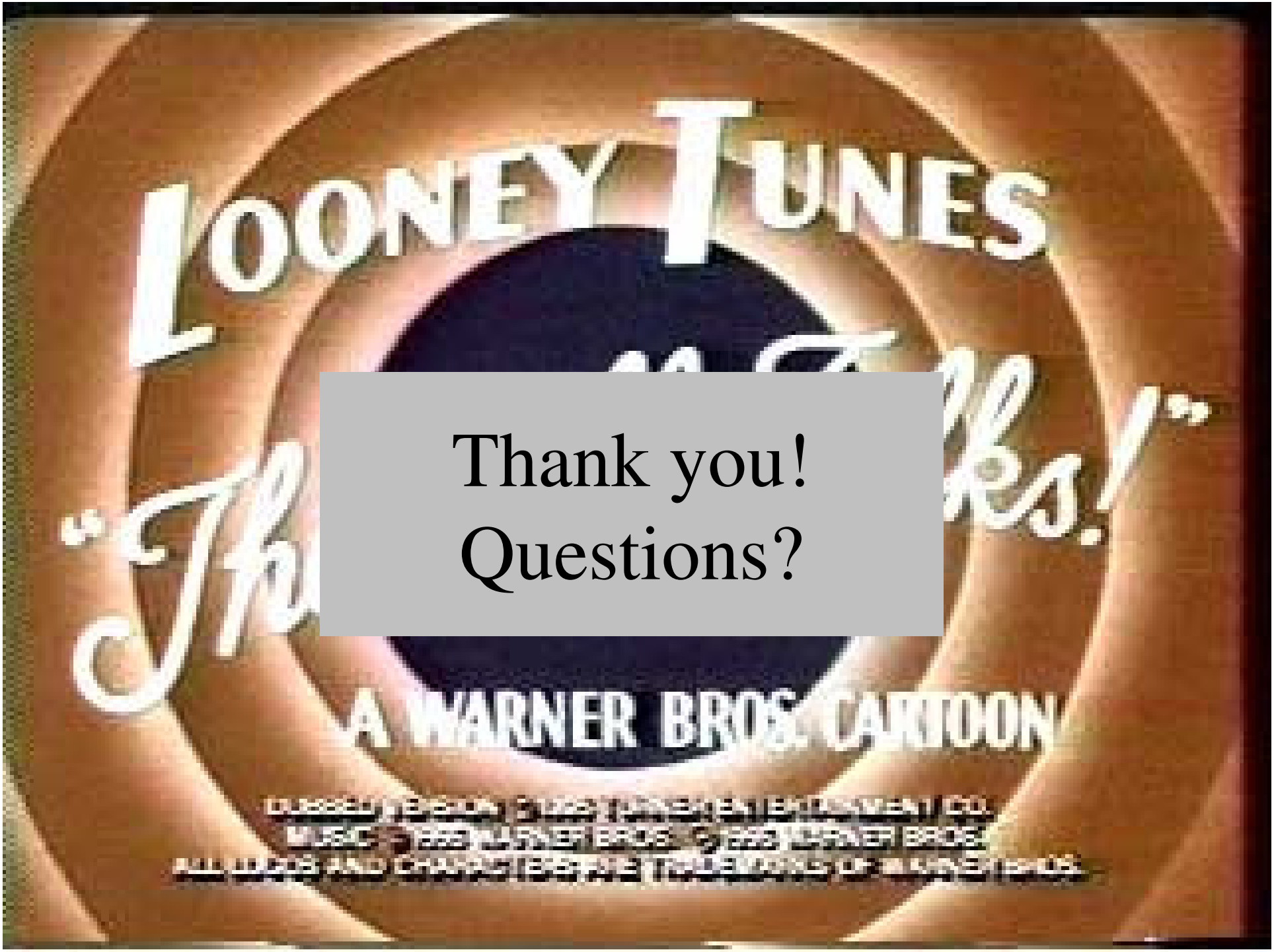
- Investigate on how to derive from the MAX-CMO model a proper similarity metric and test this metric for biological significance.

- Implement a web-server with our methodology

Future Work(2)

Goldman et.al. (GolIstPap99) present the following desiderata for a structural similarity metric:

- it should not penalize too heavily insertions and deletions ✓
- it should be reasonably robust, in that small perturbations of the definition should not make too much difference in the measure ✓
- it should be easy to compute (or at least rigorously approximated) ✓
- it should be able to discover both local and global alignments ✓
- it should be able to discover hydrophilic-hydrophobic alignments ✓
- it should take into account the self-avoiding nature of a protein ✓
- it should be subject to empirical studies on Protein Data Base (PDB) data to validate its success in capturing structural similarity ✓
- even if one comes up, from a theoretical standpoint, with a ``perfect'' measure, it will be difficult to displace entrenched measures, used for years by protein scientists. Acceptance in the field is thus a further desideratum. ✓



Thank you!
Questions?