Protein and RNA Structure Alignment

Laboratory of Bioinformatics I
Module 2

Emidio Capriotti
http://biofold.org/

Biomolecules
Folding and
Disease

Department of Pharmacy and
Biotechnology (FaBiT)
University of Bologna

ALMA MATER STUDIORUM
A.D. 1088

Structure Superimposition

Given two sets of points with some the dimension $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1,b_2,\ldots b_n)$ in Cartesian space, find the optimal rigid body transformation G between the two subsets A and B that minimizes a given distance metric D over all possible rigid body transformation G, i.e.

$$\boldsymbol{Y = G(X) = A * X + B}$$

A = 3x3 rotation matrix

B = the translation vector

X = original point

$$\mathbf{RMSD} = \sqrt{\frac{\sum_{i=1}^{n} (a_i - b_i)^2}{n}}$$

$$\mathbf{A} = \begin{bmatrix} \cos\theta\cos\psi & \cos\phi\sin\psi + \sin\phi\sin\theta\cos\psi & \sin\phi\sin\psi - \cos\phi\sin\theta\cos\psi \\ -\cos\theta\sin\psi & \cos\phi\cos\psi - \sin\phi\sin\theta\sin\psi & \sin\phi\cos\psi + \cos\phi\sin\theta\sin\psi \\ \sin\theta & -\sin\phi\cos\theta & \cos\phi\cos\theta \end{bmatrix}$$

Therefore structural superimposition correspond the best rototraslation which computational complexity is O(n).

Structural Alignment

Given two sets of points $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots b_m)$ in Cartesian space, find the optimal subsets A(P) and B(Q) with |A(P)| = |B(Q)|, and find the optimal rigid body transformation G between the two subsets A(P) and B(Q) that minimizes a given distance metric D over all possible rigid body transformation G, i.e.

$$\boxed{\min_G \left\{ D\left[A(P) - G(B(Q)) \right] \right\}}$$

$$\mathbf{RMSD} = \sqrt{\frac{\sum_{i=1}^{n} (\mathbf{a}_i - \mathbf{b}_i)^2}{\mathbf{n}}}$$

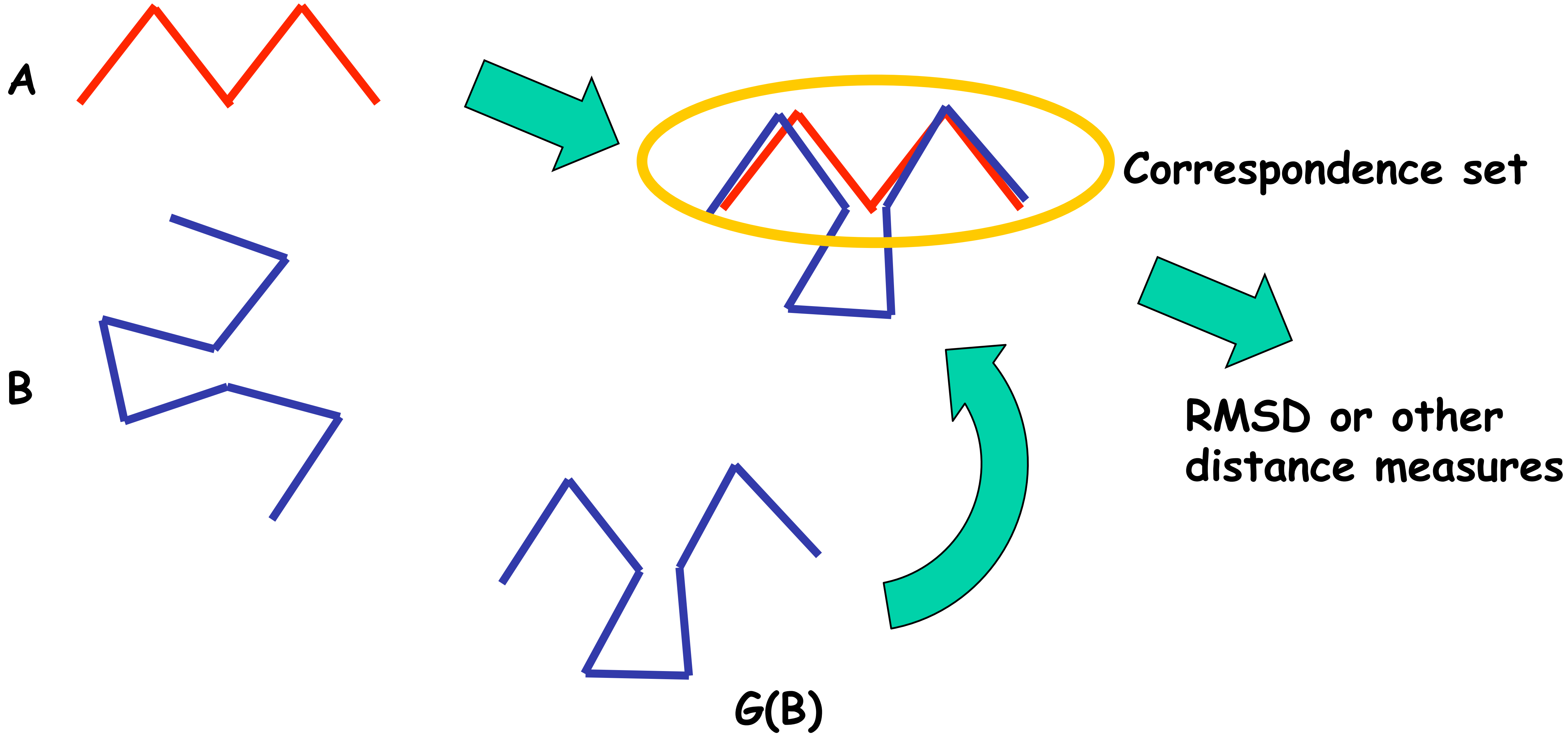
The two subsets A(P) and B(Q) define a “correspondence”, and p = |A(P)| = |B(Q)| is called the correspondence length. Naturally, the correspondence length is maximal when A(P) and B(Q) are similar.

Therefore there are essentially two problems in structure alignment:

- Find the correspondence set (which is NP-hard), and
- Find the alignment transform (which is O(n)).

Bourne P. 2012

Structural Alignment



Correspondence: (A_1,B_1), (A_2,B_2), (A_3,B_6), (A_4,B_7), (A_5,B_8)

Superimposition vs Alignment

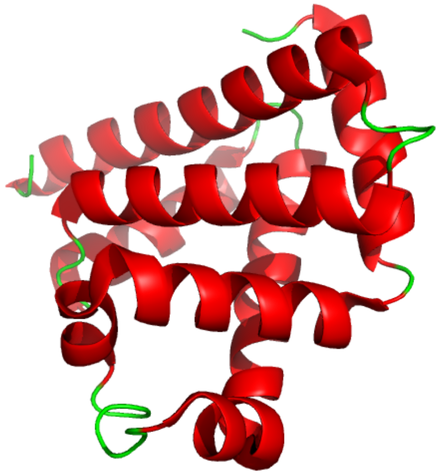
- Structure superposition assumes you already know which atoms to superimpose (correspondence set)

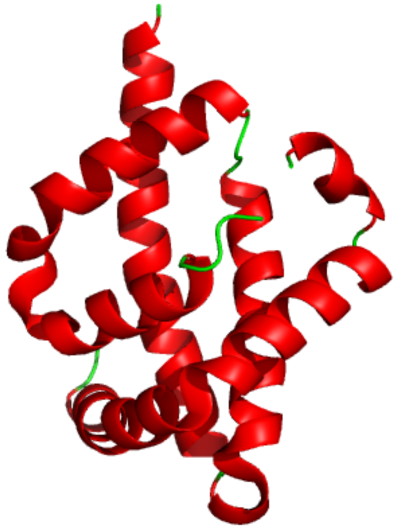
  it merely optimizes the position of the chosen atoms (relatively simple)

- Structure alignment must first determine what atoms to align (difficult).

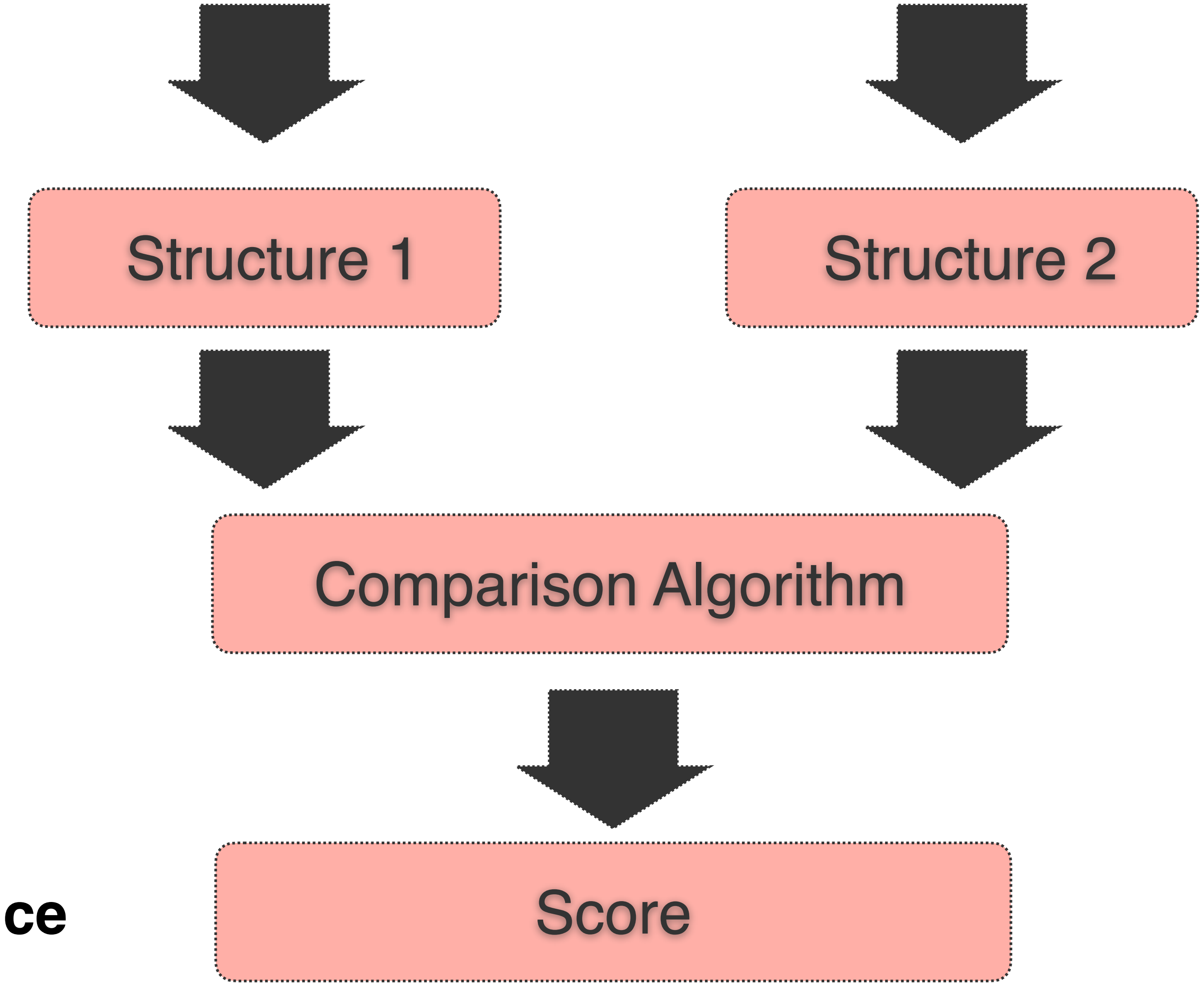
Structures Comparison

Sperm Whale
Myoglobin
(1JP6:A)





Bacterial
Haemoglobin
(1VHB:A)



Feature
Extraction

Structure 1

Structure 2

Algorithm

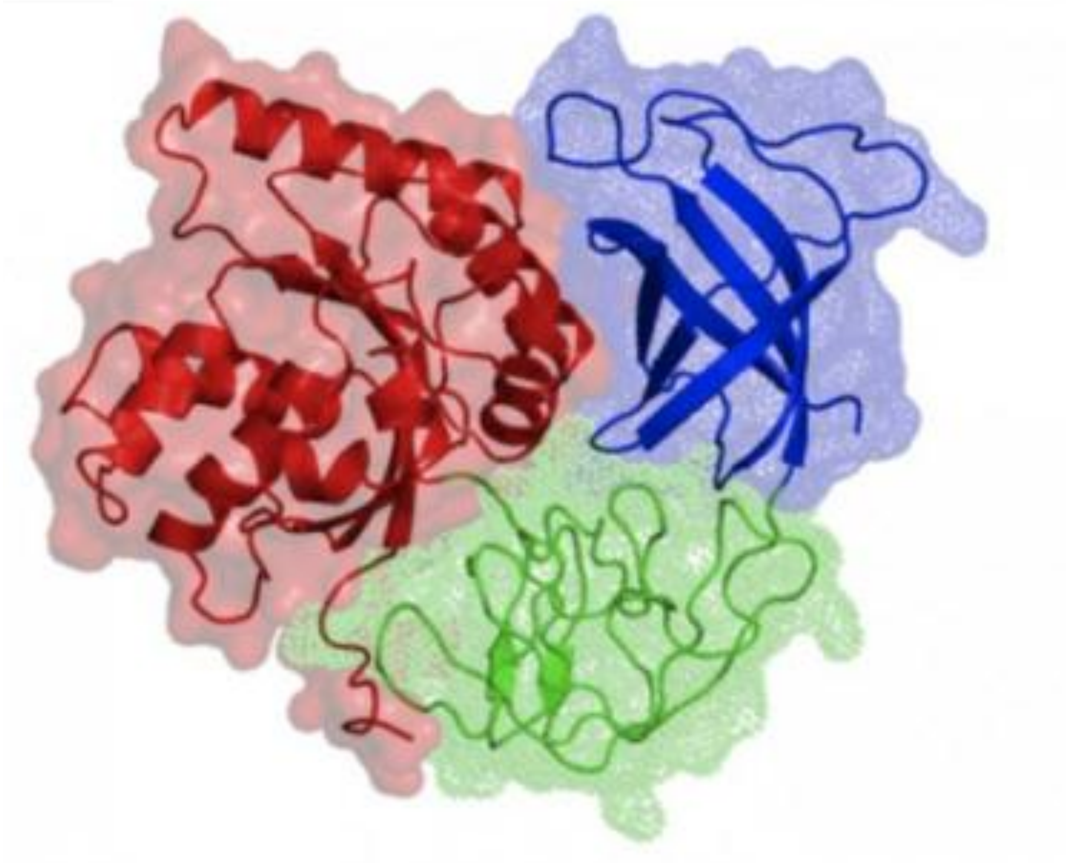
Comparison Algorithm

Statistical Significance

Score

Bourne P. 2012

Level of Comparison

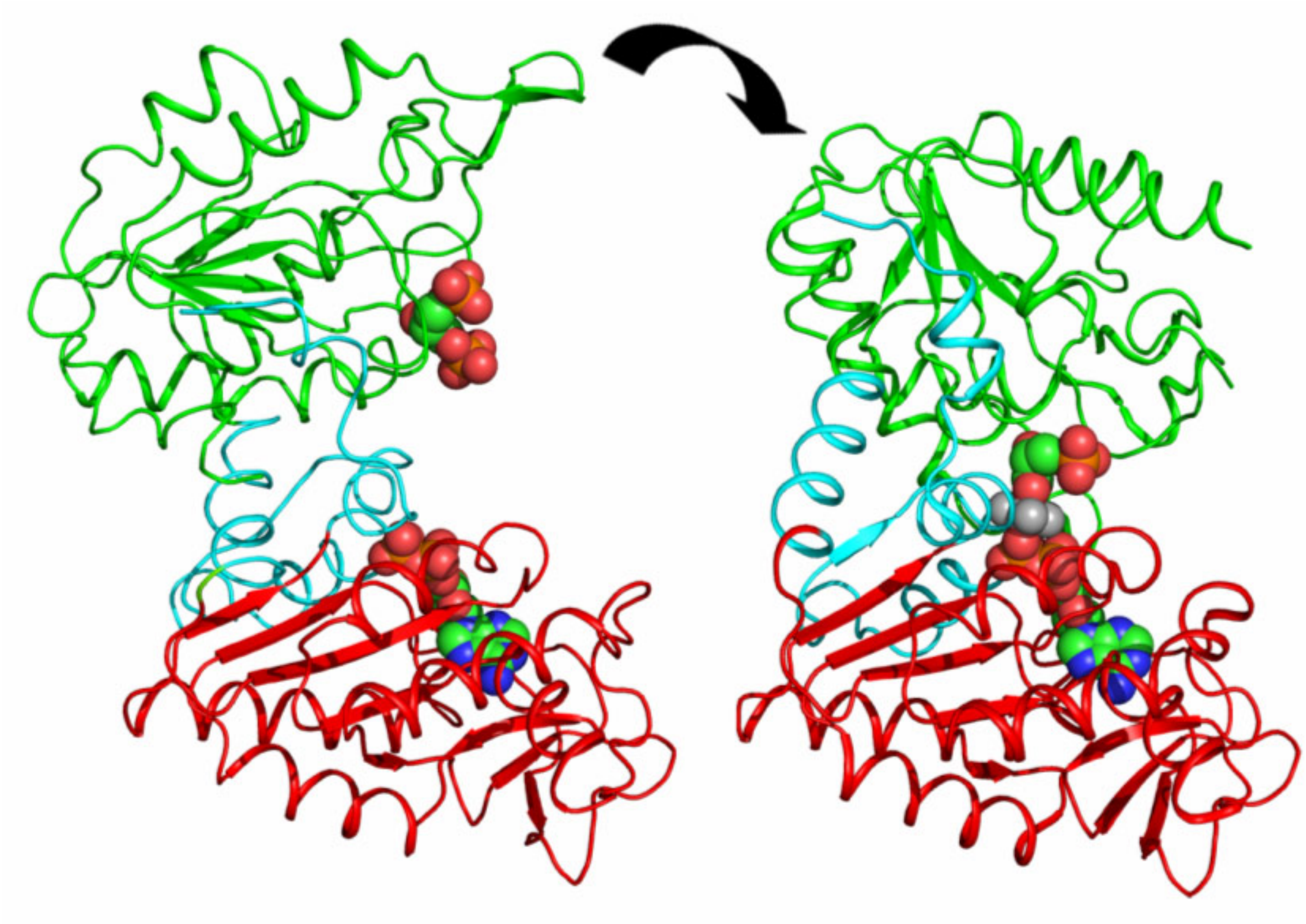


Three domains of Thermus aquaticus elongation factor EF-Tu:
in blue (all-β), red (α/β) and green (all-β).

Structural domains (the units of fold) are independently stable tertiary structures of proteins. They are distinct functional and/or structural units and can evolve, exist and function independently. Therefore, the same domain can be a part of different protein (EBI on-line course)

The definition of domain is often heuristic and questionable. The independent evolution/existence and functionality is rarely experimentally tested.

Multi Domain Alignment



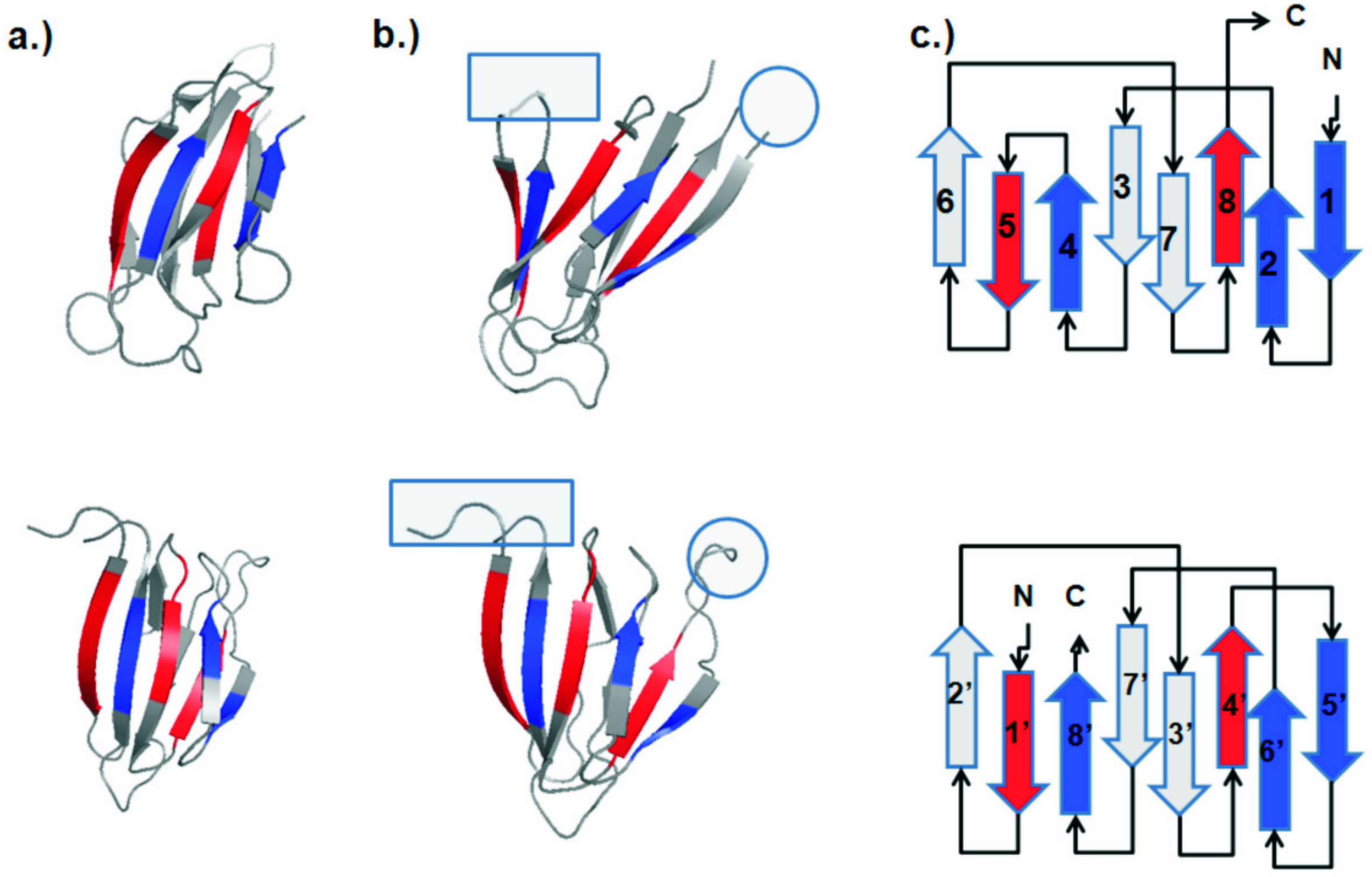
Domain movements in PGK catalysis. The fully-open resting state of the enzyme defined by refinement against SAXS data (left) binds the substrates 13BPG in the N domain (green) and ADP in the C-domain (red).

A rotation of ~56° of the hinge region (blue) brings the substrates together to initialise catalysis and ATP production (right).

Image credit: M.W. Bowler

Topology Independent Alignments

Most protein structural alignment methods can reliably classify proteins into similar folds given the structural units from each protein are in the same sequential order. However, the evolutionary possibility of proteins with different structural topology but with similar spatial arrangement of their secondary structures pose a problem.

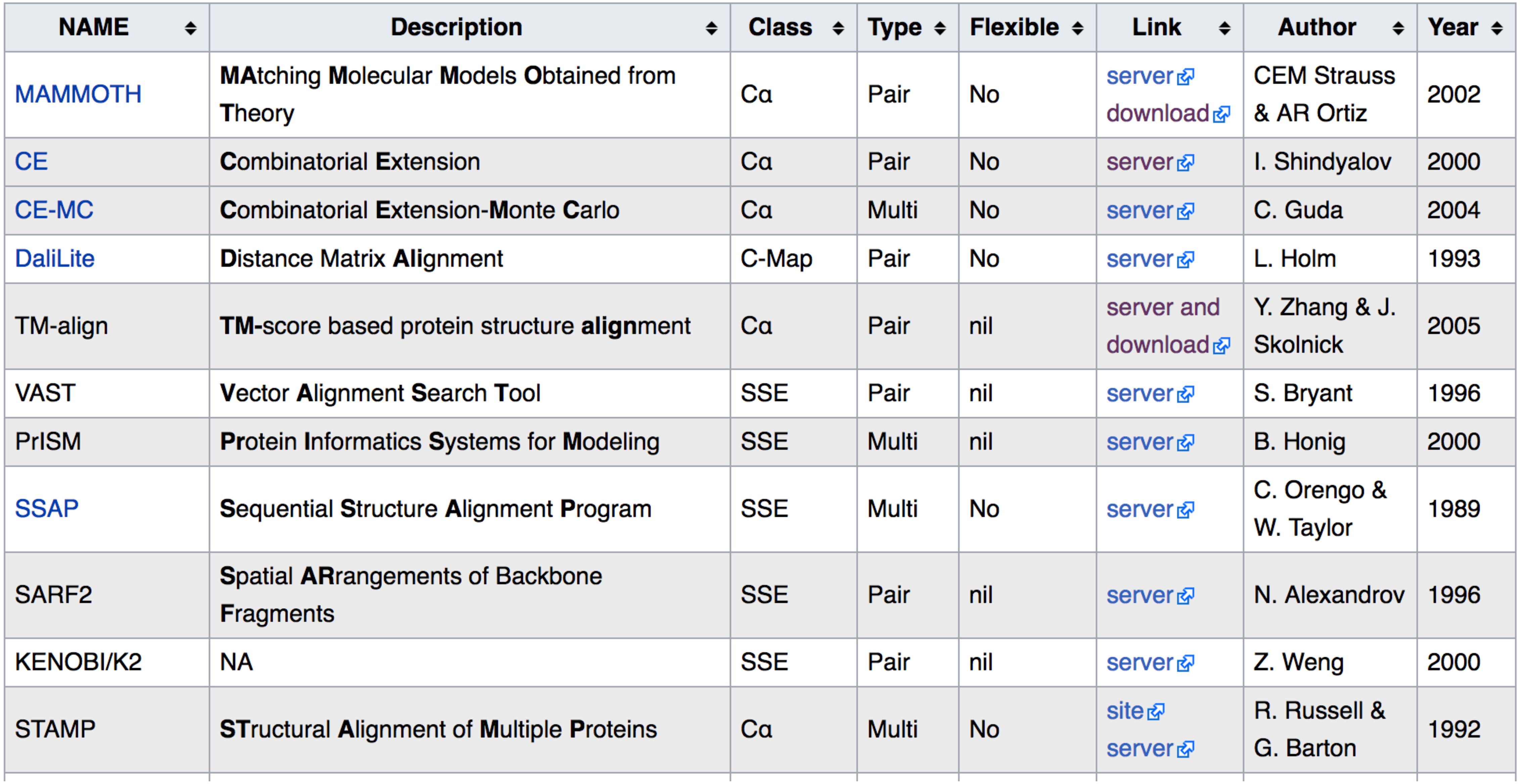


Nucleoplasmin-core (1k5j, chain E, top panel), and the fragment of residues 37–127 of auxin binding protein 1 (1lrh, chain A, bottom panel). a) These two proteins superimpose well spatially, with an RMSD value of 1.36Å for an alignment length of 68 residues.

Dundas et al. (2007) PMID:17937816

Structural Alignment Tools

There are several well-documented, easy to use software packages for structural alignment. More than 100 are reported on wikipedia.



NAME	Description	Class	Type	Flexible	Link	Author	Year
MAMMOTH	**MA**tching **M**olecular **M**odels **O**btained from **T**heory	Cα	Pair	No	server download	CEM Strauss & AR Ortiz	2002
CE	**C**ombinatorial **E**xtension	Cα	Pair	No	server	I. Shindyalov	2000
CE-MC	**C**ombinatorial **E**xtension-**M**onte **C**arlo	Cα	Multi	No	server	C. Guda	2004
DaliLite	**D**istance Matrix **Ali**gnment	C-Map	Pair	No	server	L. Holm	1993
TM-align	**TM-**score based protein structure **align**ment	Cα	Pair	nil	server and download	Y. Zhang & J. Skolnick	2005
VAST	**V**ector **A**lignment **S**earch **T**ool	SSE	Pair	nil	server	S. Bryant	1996
PrISM	**P**rotein **I**nformatics **S**ystems for **M**odeling	SSE	Multi	nil	server	B. Honig	2000
SSAP	**S**equential **S**tructure **A**lignment **P**rogram	SSE	Multi	No	server	C. Orengo & W. Taylor	1989
SARF2	**S**patial **AR**rangements of Backbone **F**ragments	SSE	Pair	nil	server	N. Alexandrov	1996
KENOBI/K2	NA	SSE	Pair	nil	server	Z. Weng	2000
STAMP	**ST**ructural **A**lignment of **M**ultiple **P**roteins	Cα	Multi	No	site server	R. Russell & G. Barton	1992

https://en.wikipedia.org/wiki/Structural_alignment_software

Method Classification

Type

Pair Pairwise Alignment (2 structures only);

Multi Multiple Structure Alignment;

Class

Cα Backbone Atom (Cα) Alignment;

AllA All Atoms Alignment;

SSE Secondary Structure Elements Alignment;

Seq Sequence-based alignment

C-Map Contact Map

Surf Connolly Molecular Surface Alignment

SASA Solvent Accessible Surface Area

Dihed Dihedral Backbone Angles

PB Protein Blocks

Protein descriptors

Flexible

No Only rigid-body transformations are considered between the structures being compared.

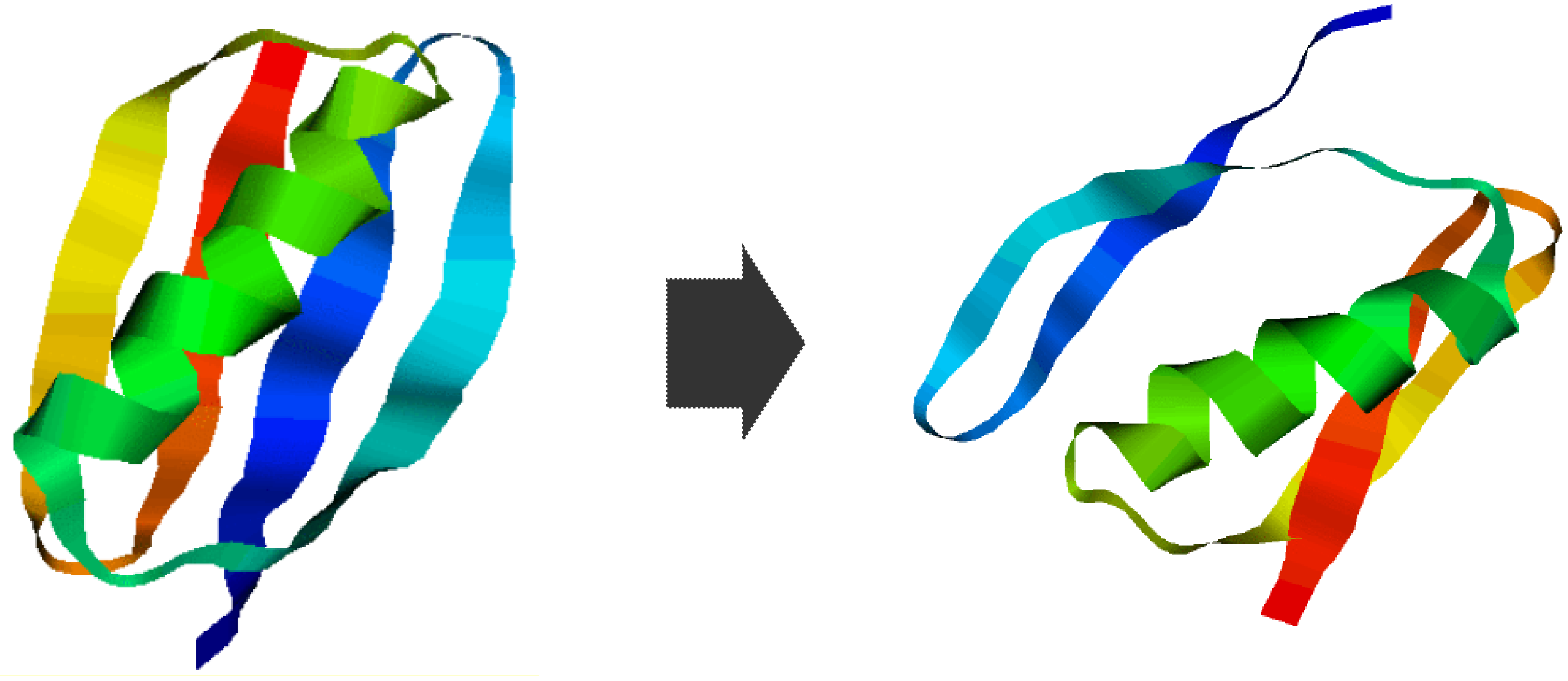
Yes The method allows for some flexibility within the structures being compared, such as movements around hinge regions.

Comparing Torsion Angles

Torsion Angles (Φ,Ψ) are:

- local by nature
- invariant upon rotation and translation of the molecule
- compact - complexity o(n)

Good for alignment of local region but
possible problems on the alignment of the whole structure.



Credit: Predrag Radivojac

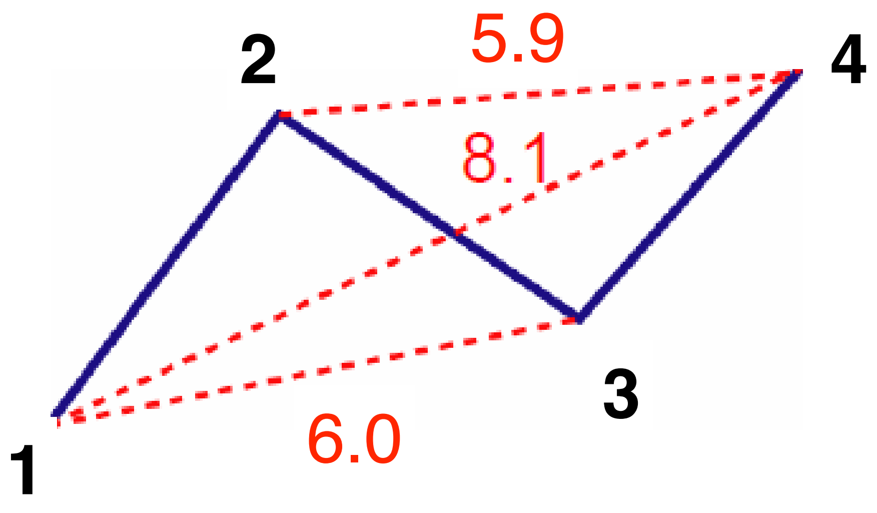
Distance Matrix

Advantage:

- invariant upon rotation and translation of the molecule
- can be used for protein comparison

Disadvantages

- Comparing matrices is an hard computational problem
- Complexity is o(n^2) where n represents the number of residues
- Insensitive to chirality



	1	2	3	4
1	0.0	3.8	6.0	8.1
2	3.8	0.0	3.8	5.9
3	6.0	3.8	0.0	3.8
4	8.1	5.9	3.8	0.0

Credit: Predrag Radivojac

Structural Alignment Components

Input & output of alignment algorithm

Input: two proteins: $A = \{a_1, \cdots, a_m\}$ $B = \{b_1, \cdots, b_n\}$

Output: An alignment and scores

$$L(A,B) = \{(a_{i_1}, b_{j_1}), \cdots, (a_{i_L}, b_{j_L})\},$$

$$i_1 < i_2 < \cdots < i_L, j_1 < j_2 < \cdots < j_L$$

Constraints:

min rmsd:

$$rmsd = \min_T \sqrt{\frac{\sum_{k=1}^{L} (a_{i_k} - Tb_{j_k})^2}{L}}$$

max L

min Gaps

$$Gaps = \sum_{t=1}^{L-1} \left[\left(i_{t+1} - i_t - 1\right) + \left(j_{t+1} - j_t - 1\right) \right]$$

Dynamic programming, Integer programming, Monte Carlo…

Statistical Significance

Phil Bourne 2012

State of the art

- All methods can identify obvious similarities between two structures
- Remote similarities are detected by a subset of methods – different remote similarities are recognized by different methods
- Good alignments are much harder to come by
- Speed is a serious issue with some algorithms

Phil Bourne 2012

Desirable Method Features

- Biologically meaningful alignments not just geometrically meaningful

- Complete database of all alignments

- Ability to apply to structures not in the PDB

Phil Bourne 2012

CE Algorithm

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

Shindyalov and Bourne (1998) PMID 9796821

Constrain the search

The alignment between two proteins A and B is the longest continuous path P of AFPs of size m in a similarity matrix

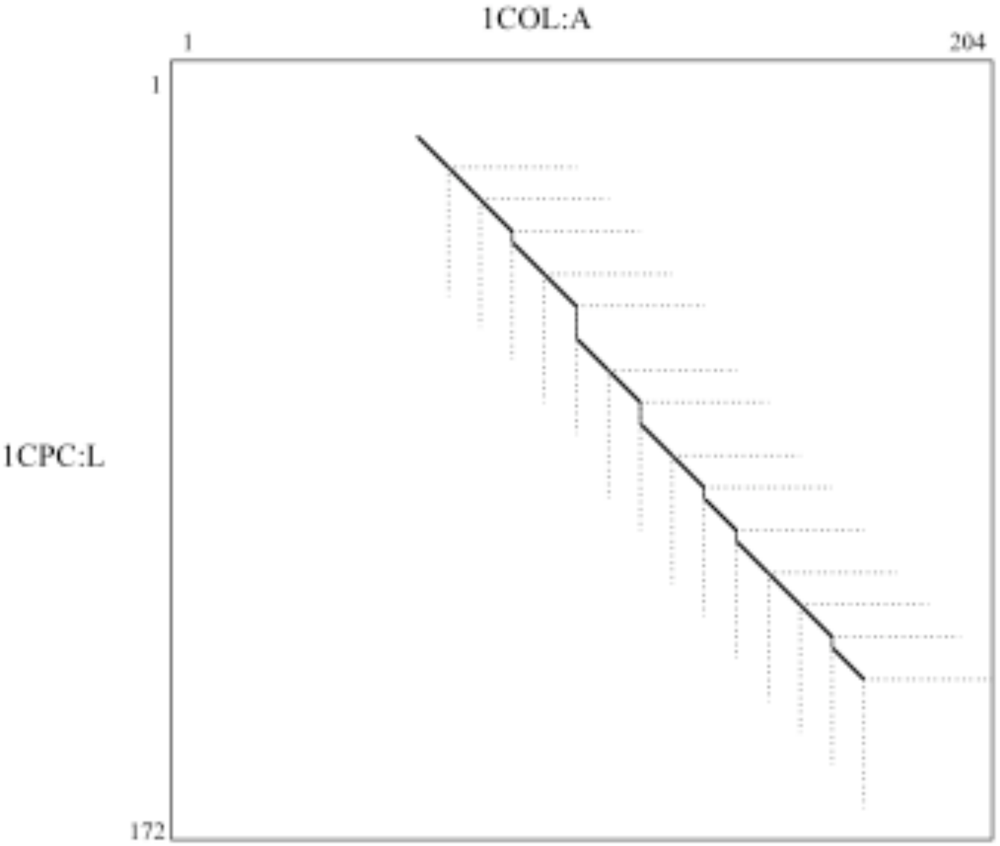
Similarity Matrix S represents all AFPs conforming to some similarity criterion (e.g., low RMSD):

$$S=(n_A-m+1)\times(n_B-m+1)$$

m = Length of AFP

n_A = Length of protein A

n_B = Length of protein B



This is very large to compute – constraints are needed

Shindyalov and Bourne (1998) PMID 9796821

Path Definition

$p^A{}_i$ = AFPs starting residue position in protein A at the i-th position of the alignment path

m = longest continual path – set as 8

One of the conditions (1)-(3) should be satisfied for 2 consecutive AFPs i and i+1 in the path

(1) = 2 consecutive AFPs aligned without gaps
(2) = Two consecutive AFPs with a gap in protein A
(3) = Two consecutive AFPs with a gap in protein B

$$p^A_{i+1} = p^A_i + m \text{ and } p^B_{i+1} = p^B_i + m \qquad (1)$$

or

$$p^A_{i+1} > p^A_i + m \text{ and } p^B_{i+1} = p^B_i + m \qquad (2)$$

or

$$p^A_{i+1} = p^A_i + m \text{ and } p^B_{i+1} > p^B_i + m \qquad (3)$$

Phil Bourne 2012

Extension of the Path

Gap sizes are limited to G – heuristically set as 30 residues

$$p_{i+1}^{A} \leqslant p_{i}^{A} + m + G \tag{4}$$

$$p_{i+1}^{B} \leqslant p_{i}^{B} + m + G \tag{5}$$

Similarity Measures

1. RMSD from least squares superposition used to select few best fragments

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

$$D_{ij} = \frac{1}{m^2}\left(\sum_{k=0}^{m-1} \sum_{l=0}^{m-1} \left| \; d^A_{p_i^A+k, p_j^A+l} - d^B_{p_i^B+k, p_j^B+l} \; \right| \right) \qquad (7)$$

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

$$D_{ij} = \frac{1}{m}\left(\left| d^A_{p_i^A p_i^A} - d^B_{p_i^B p_i^B} \right| + \left| d^A_{p_i^A+m-1, p_j^A+m-1} - d^B_{p_i^B+m-1, p_j^B+m-1} \right| + \sum_{k=1}^{m-2} \left| \; d^A_{p_i^A+k, p_j^A+m-l-k} - d^B_{p_i^B+k, p_j^B+m-l-k} \right| \right) \qquad (6)$$

Phil Bourne 2012

Statistical Evaluation

Evaluate the probability of finding an alignment path of the same length or smaller gaps and distance from a random set of non-redundant structures.

Optimization:

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

Phil Bourne 2012

Limitations

- Will not find non-topological alignments (outside the bounds of the dotted lines)
- What are the correct “units” to be comparing?
- CE initially worked on chains – as we shall see in future weeks domains are the correct units, but definition of the domains is not straightforward

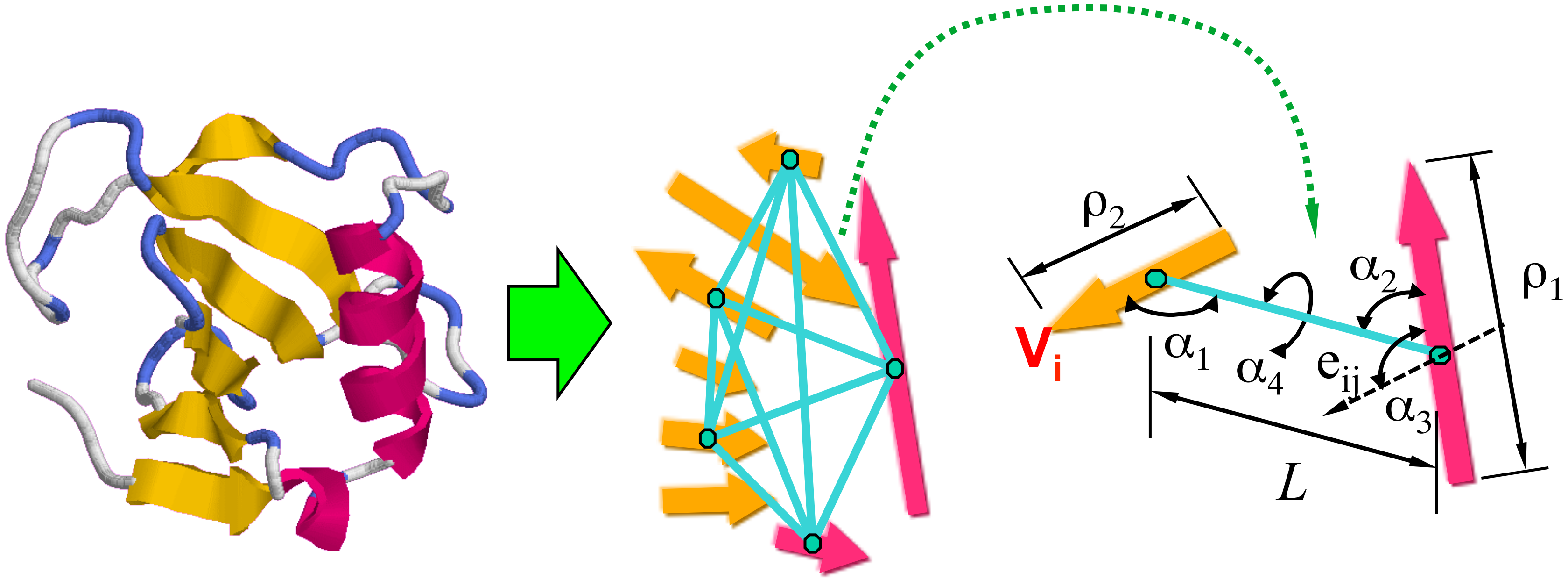
Phil Bourne 2012

PDBe Fold

- Protein secondary structure elements (SSE) – natural and convenient objects for building three dimensional graphs.
- Secondary structures provide most functionality and is conserved through evolution
- Details of protein fold – expressed in terms of two SSE – helices and strands

Phil Bourne 2012

Graph Representation (I)



SSE graphs- represented by vectors

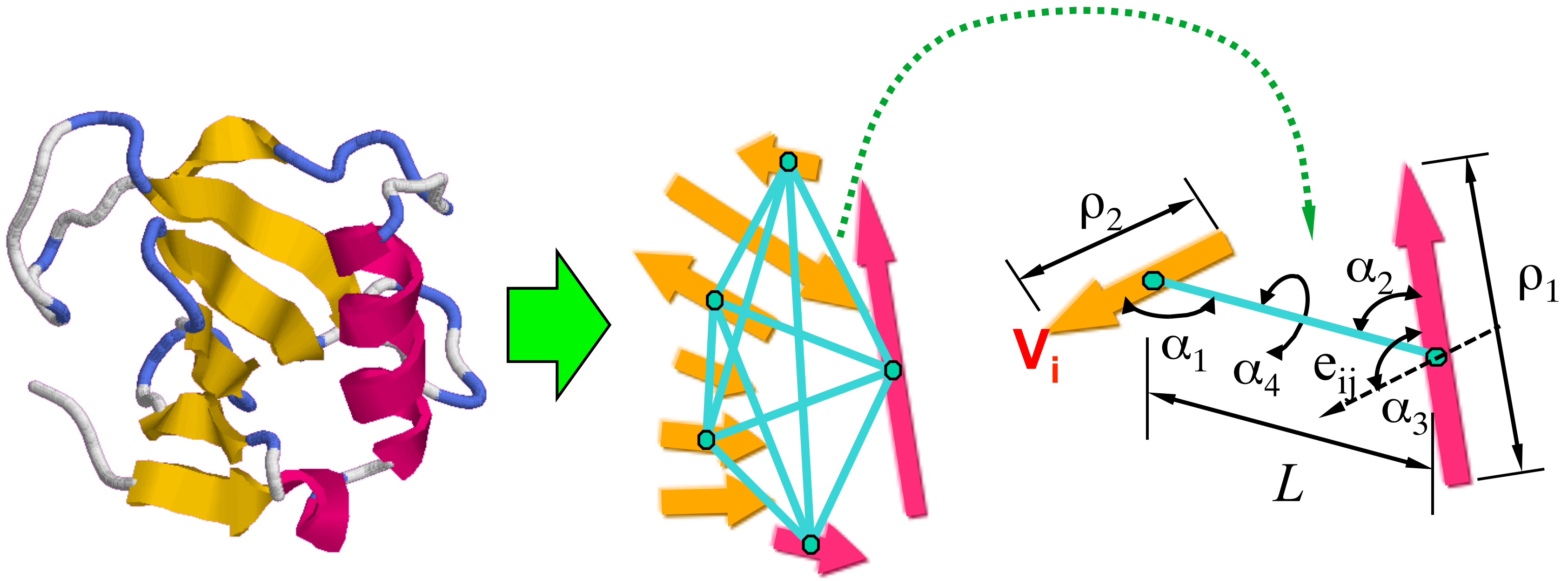
Each SSE can be used as graph vertices (T_i, ρ_i)

Any 2 vertices are connected by an edge label L – describes position and orientation of the connected SSEs

Each edge labelled with a property vector – $\alpha_{1/2}$ angle between edge and vertices, torsion angle between vertices, length of the edge L

Krissinel and Henrick (2004) PMID: 15572779

Graph Representation (II)



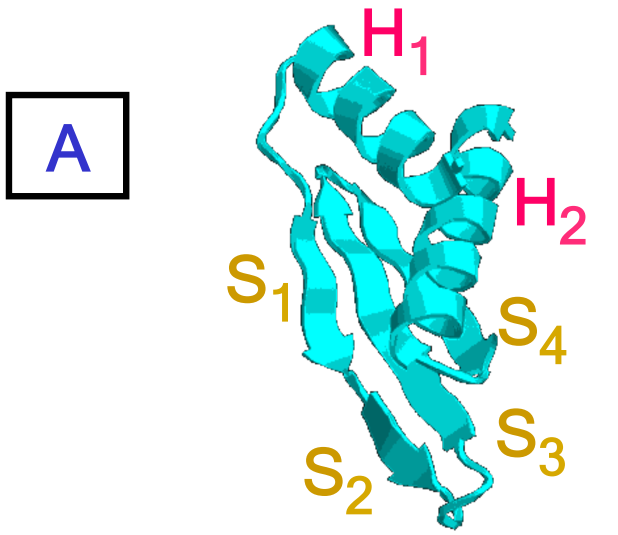
Sets of vertices, edges and their labels provides full definition of the graph.

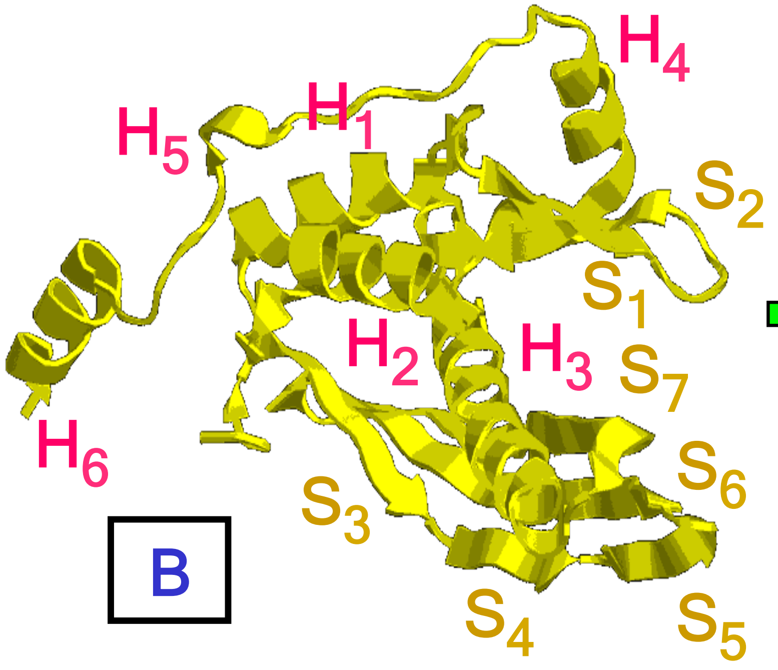
Graph matching algorithm is required – set of rules for comparing individual vertices and edges – tolerances chosen empirically

Relative and absolute vertex and edge lengths are used for comparison – allows larger absolute differences for longer vertices and edges

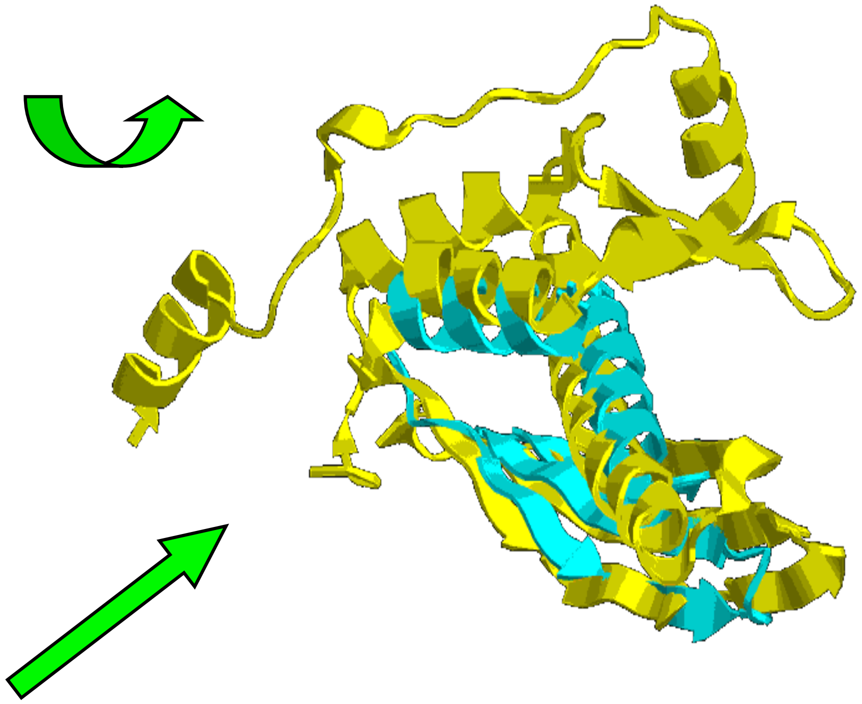
Torsion angle comparison – distinguish mirror symmetry mates

Graph Matching





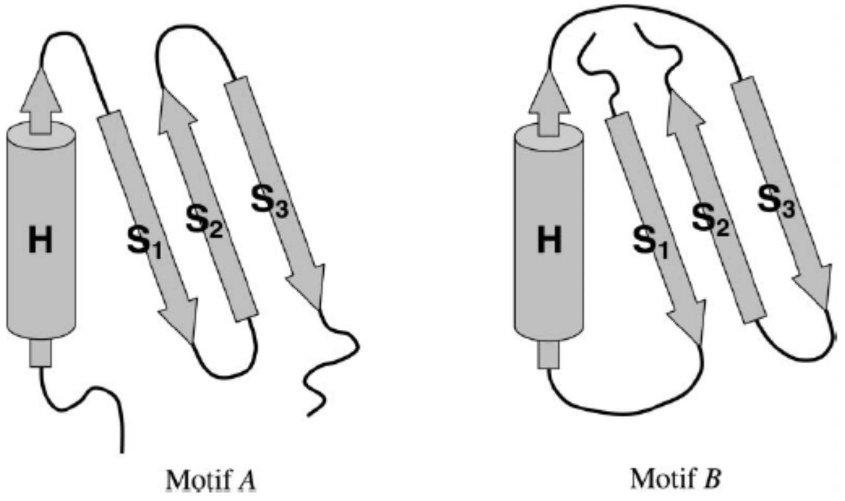
A	B
	H_1
	S_1
	S_2
H_1	H_2
S_1	S_3
S_2	S_4
	S_5
S_3	S_6
S_4	S_7
H_2	H_3
	H_4
	H_5
	H_6



Matching the SSE graphs yields a correspondence between secondary structure elements, that is, groups of residues. The correspondence may be used as initial guess for structure superposition and alignment of individual residues.

PDBe Fold Approaches

1) Connectivity of SSE Neglected



2) Soft connectivity – general order of SSEs along their protein chains are same in both structures BUT any number of missing/unmatched SSE between matched ones allowed

3) Strict connectivity – matched SSEs follow same order along their protein chains – separated only by equal number of matched/ unmatched SSE in both structures

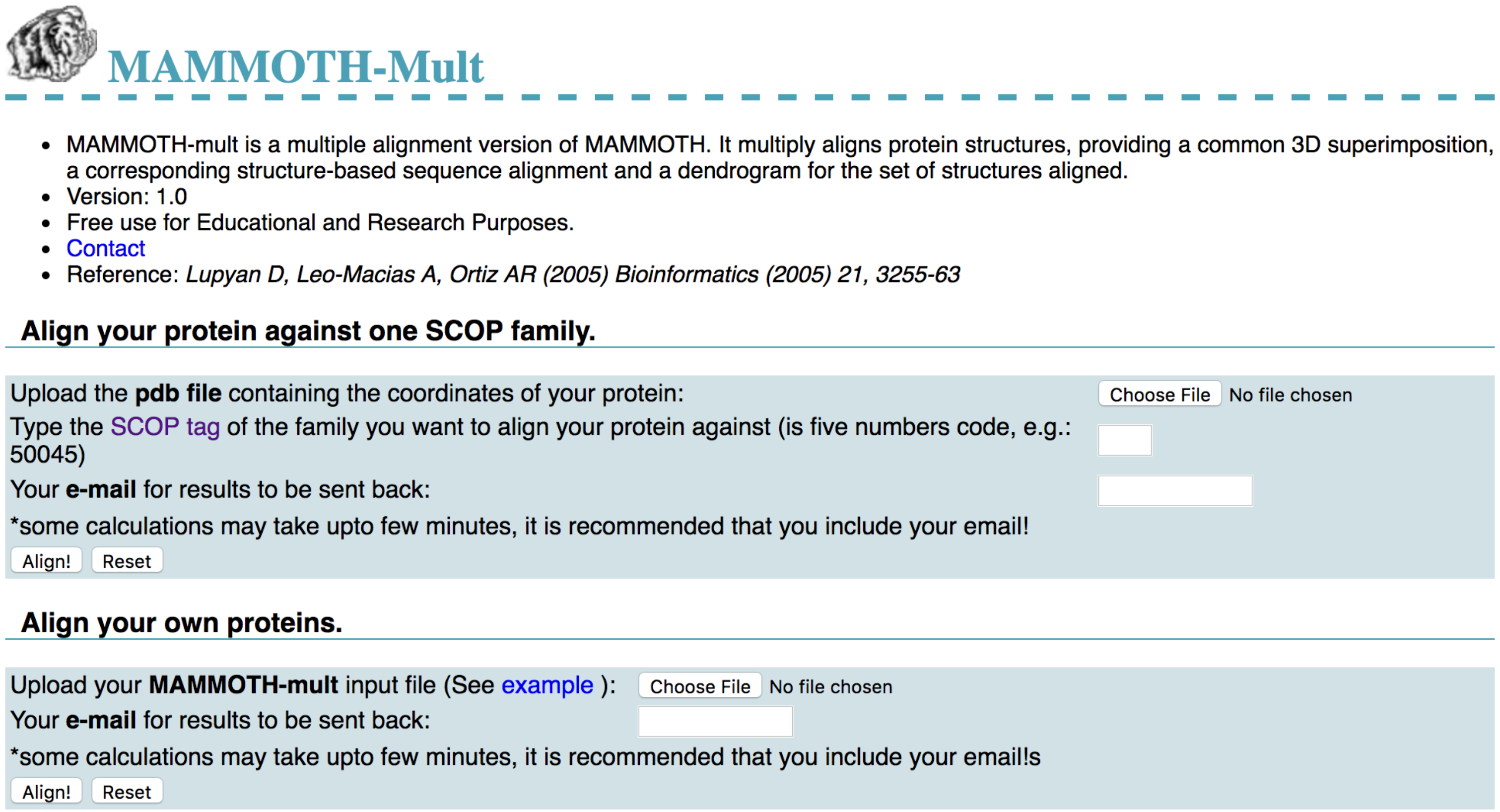
To obtain 3D alignment of individual residues – represent them by their C-alpha atoms – use results of graph matching as a starting point

MAMMOTH Algorithm

The MAMMOTH (MAtching Molecular Models Obtained from Theory) algorithm is one of the fastest methods for structural alignment .

The method represents a protein structure as a set of unit vectors build using the vectors between C-α atoms.

MAMMOTH uses a dynamic programming algorithm to find the bast alignment between two protein structure.



MAMMOTH-Mult

- MAMMOTH-mult is a multiple alignment version of MAMMOTH. It multiply aligns protein structures, providing a common 3D superimposition, a corresponding structure-based sequence alignment and a dendrogram for the set of structures aligned.
- Version: 1.0
- Free use for Educational and Research Purposes.
- Contact
- Reference: *Lupyan D, Leo-Macias A, Ortiz AR (2005) Bioinformatics (2005) 21, 3255-63*

Align your protein against one SCOP family.

Upload the **pdb file** containing the coordinates of your protein:

Type the SCOP tag of the family you want to align your protein against (is five numbers code, e.g.: 50045)

Your **e-mail** for results to be sent back:

*some calculations may take upto few minutes, it is recommended that you include your email!

Align your own proteins.

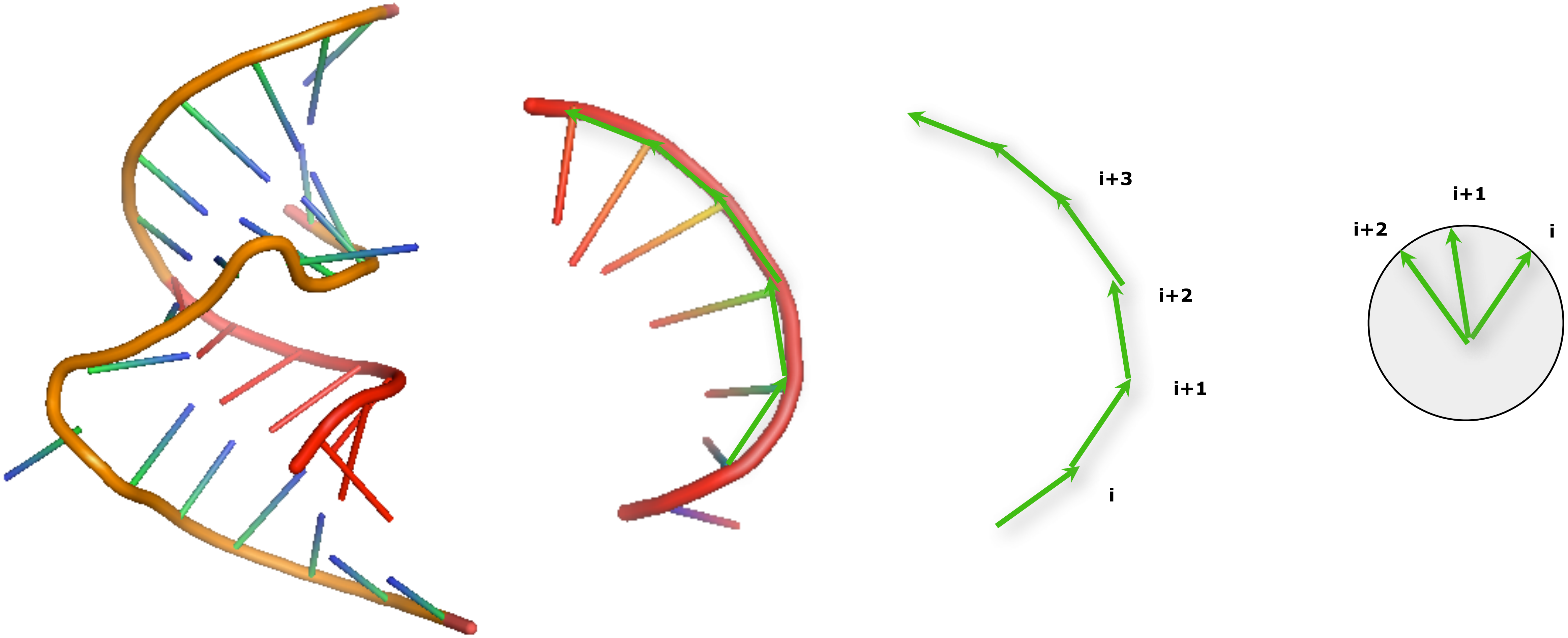
Upload your **MAMMOTH-mult** input file (See example):

Your **e-mail** for results to be sent back:

*some calculations may take upto few minutes, it is recommended that you include your emaills

https://ub.cbm.uam.es/software/online/mamothmult.php

Unit Vector Representation

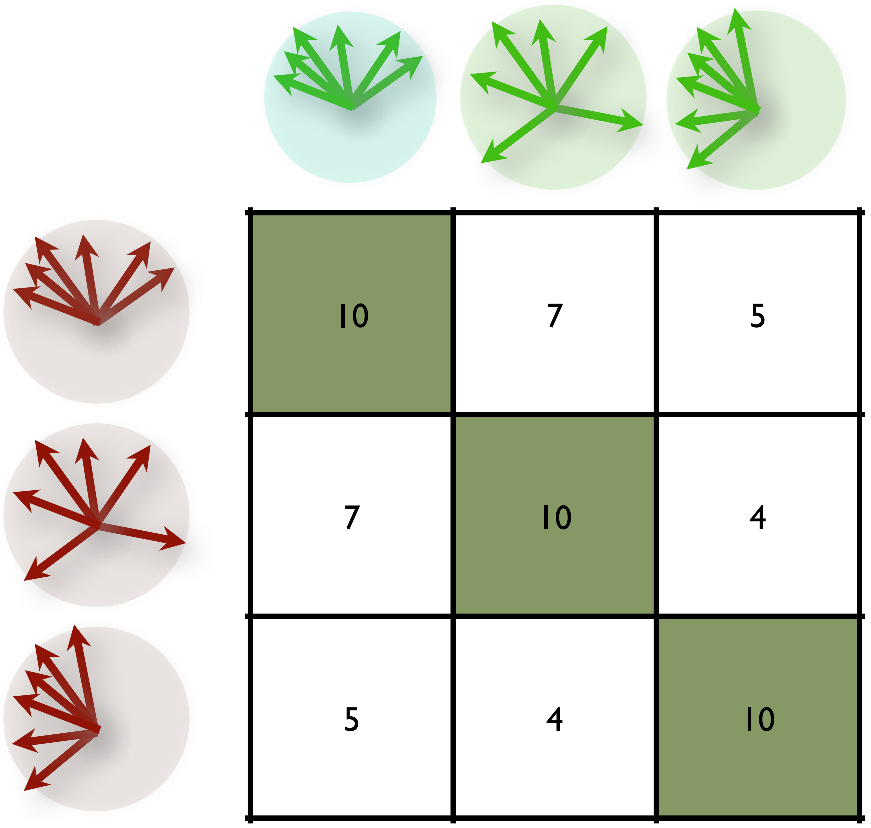


A Unit Vector is the normalized vector between two successive Cα atoms.

For each position *i* consider the *k* consecutive vectors, which will be mapped into a unit sphere representing the local structure of k residues.

Ortiz et al. (2002) PMID:12381844

Unit Vector Scoring



10	7	5
7	10	4
5	4	10

$$URMS^R = \sqrt{2.0 - \frac{2.84}{\sqrt{k}}}$$

$$S_{ij} = \frac{(URMS^R - URMS^{ij})}{URMS^R} \Delta(URMS^R, URMS^{ij})$$

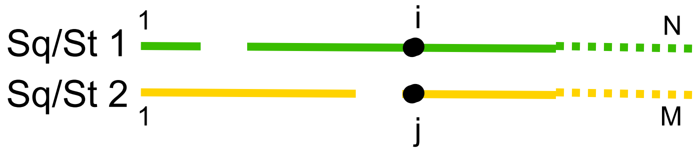
$$\Delta(URMS^R, URMS^{ij}) = 10 \Rightarrow URMS^R > URMS^{ij}$$
$$\Delta(URMS^R, URMS^{ij}) = 0 \Rightarrow URMS^R \leq URMS^{ij}$$

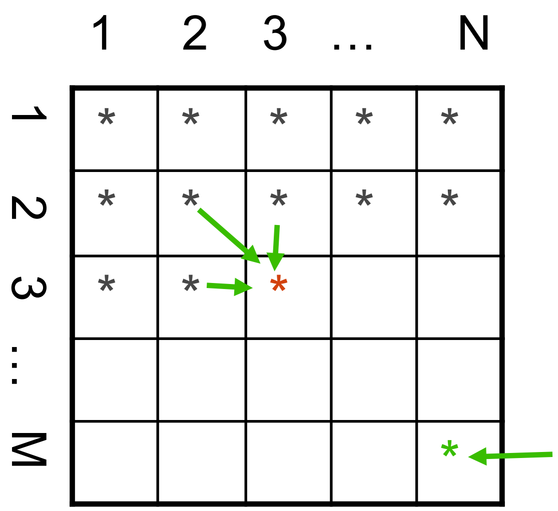
For each position i, the k consecutive unit vectors (k=6) are grouped and aligned to the j set of unit vectors. Each pair of aligned unit vectors will be evaluated by calculating Unit Root Mean Square distance (URMS^{ij}).

The obtained URMS values are compared the minimum expected URMS distance between two random set of k unit vectors (URMS^{R}).

The alignment score is than calculated normalizing URMS^{ij} to the URMS^{R} value.

Alignment





$$D_{i,j} = \min \begin{cases} D_{i,j-1} + \text{Score}_{(\text{Ä},rj)} \\ D_{i-1,j-1} + \text{Score}_{(ri,rj)} \\ D_{i-1,j} + \text{Score}_{(ri,\text{Ä})} \end{cases}$$

Backtracking to get the best alignment

A Dynamic Programming procedure is then applied to search for the optimal structural alignment using a global alignment with zero end gap penalties.

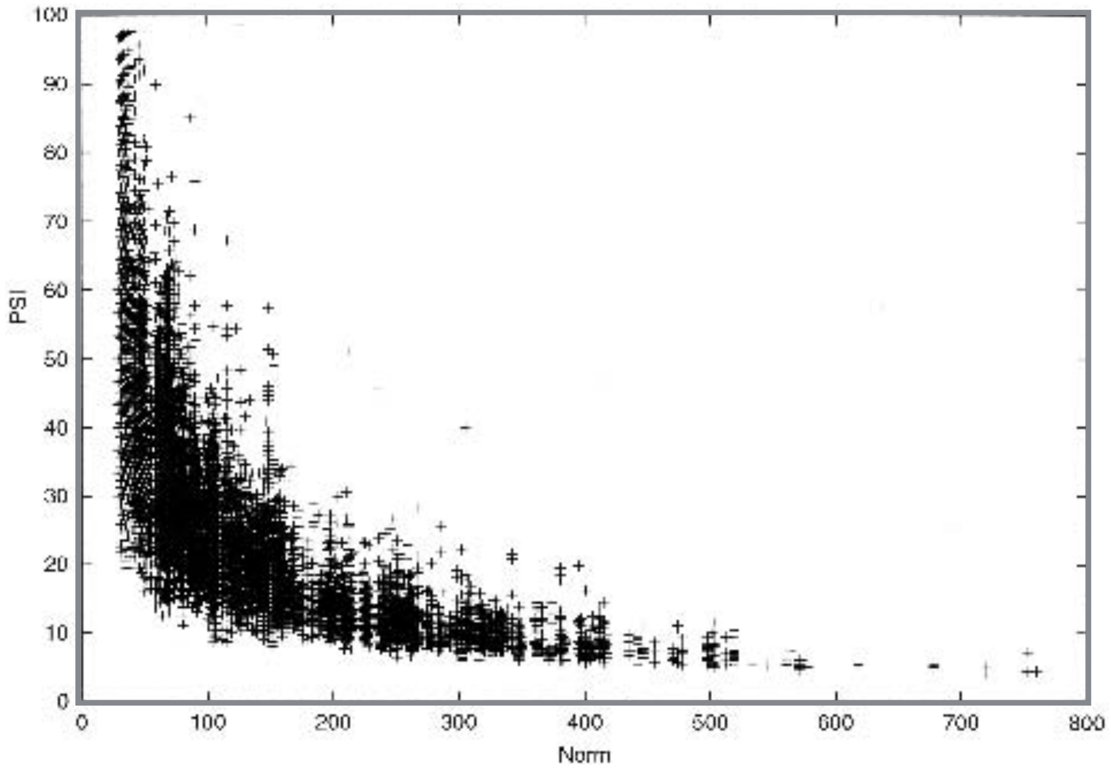
The maximum subset of local structures that have their corresponding Cα within 4.0 Å in the space are evaluated. The number of close atoms is used to evaluate the percentage of structural identity (PSI) using a variant of the MaxSub algorithm.

Siew et al. (2000) PMID: 11108700

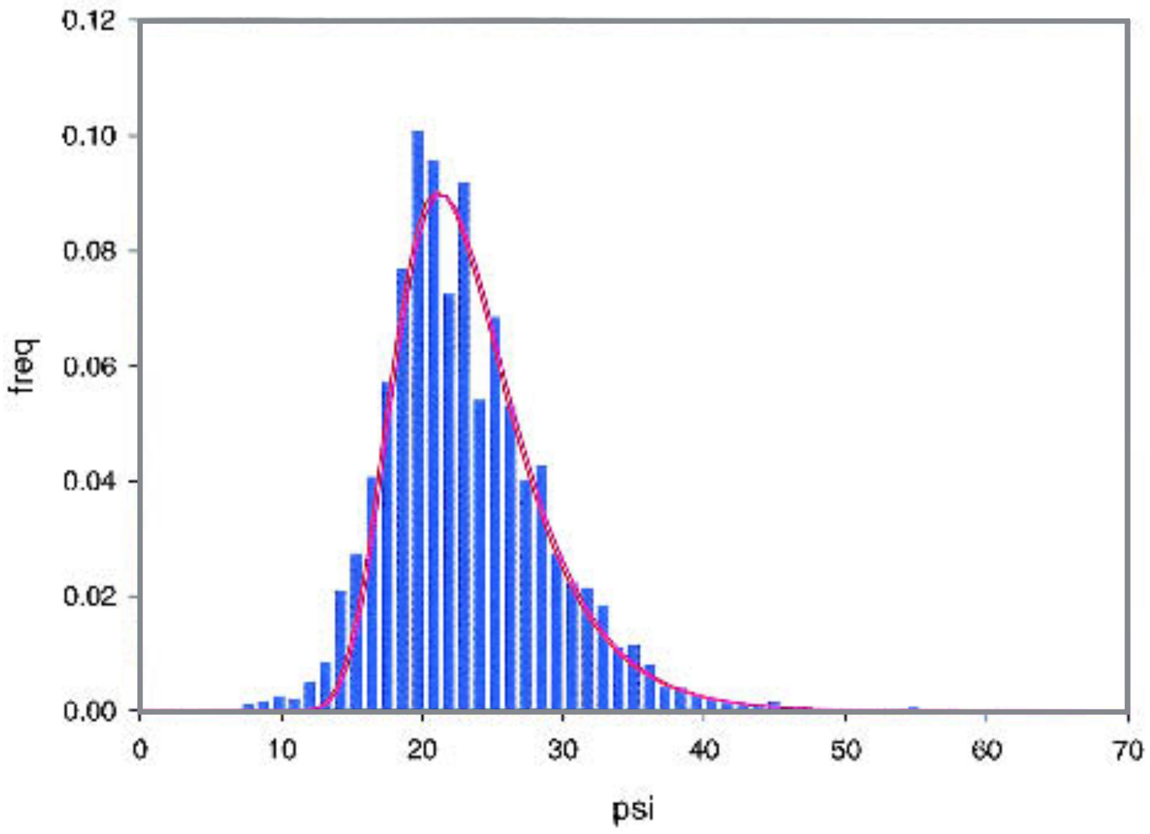
Background Distribution

Considering a dataset of random structures, it is possible to produce pairwise alignments that resulted in a empirical distribution of scores (s). From such distribution we can then evaluate μ and σ needed to calculated the p-value for P(s>x).

Empirical



Analytic



$$P(t > x) = \int_t^\infty f(x)dx = 1 - e^{-e^{\frac{-(x-a)}{b}}}$$

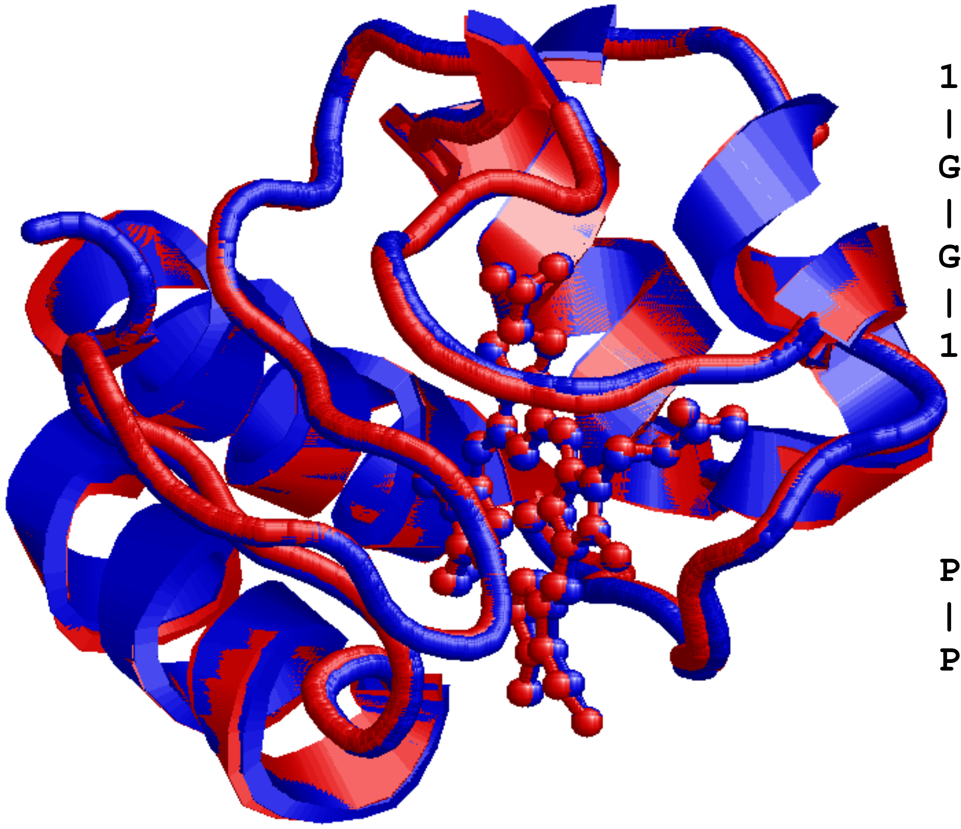
Ortiz et al. (2002) PMID:12381844

Exercise

Build a Python script for structure superimposition using the class SVDSuperimposer from the biopython libraries.

Test the script on a group of atoms from the following structures

Human Cytochrome C – Uniprot:P99999. PDB: 3ZCF:A
Equine Cytochrome C – Uniprot: P00004. PDB 3O20:A



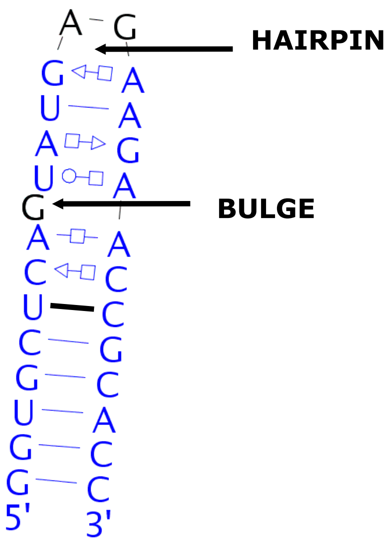
```
1:A                20:A                40:A                60:A
|        |    .    |    .    |    .    |    .    |    .    |    .    |
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLEN
||||||||||:.||.||||||||||||||||||||||||||||||:.||.|||||||.|.|:||||||||
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEETLMEYLEN
|        |    .    |    .    |    .    |    .    |    .    |    .    |
1:A                20:A                40:A                60:A


         80:A                100:A
    .    |    .    |    .    |
PKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
||||||||||||.|||||.||.||||||||||||
PKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE
    .    |    .    |    .    |
         80:A                100:A
```

RNA structure

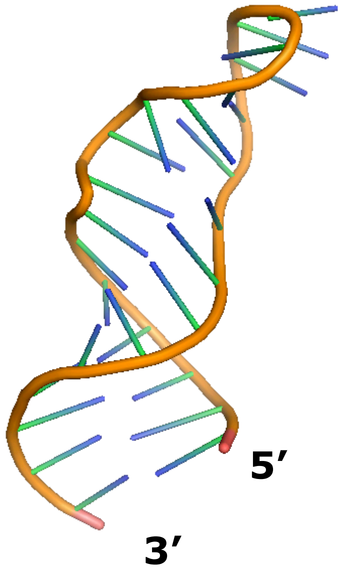
Primary Structure

>Mutant Rat 28S rRNA sarcin/ricin domain
GGUGCUCAGUAUGAGAAGAACCGCACC



Secondary Structure

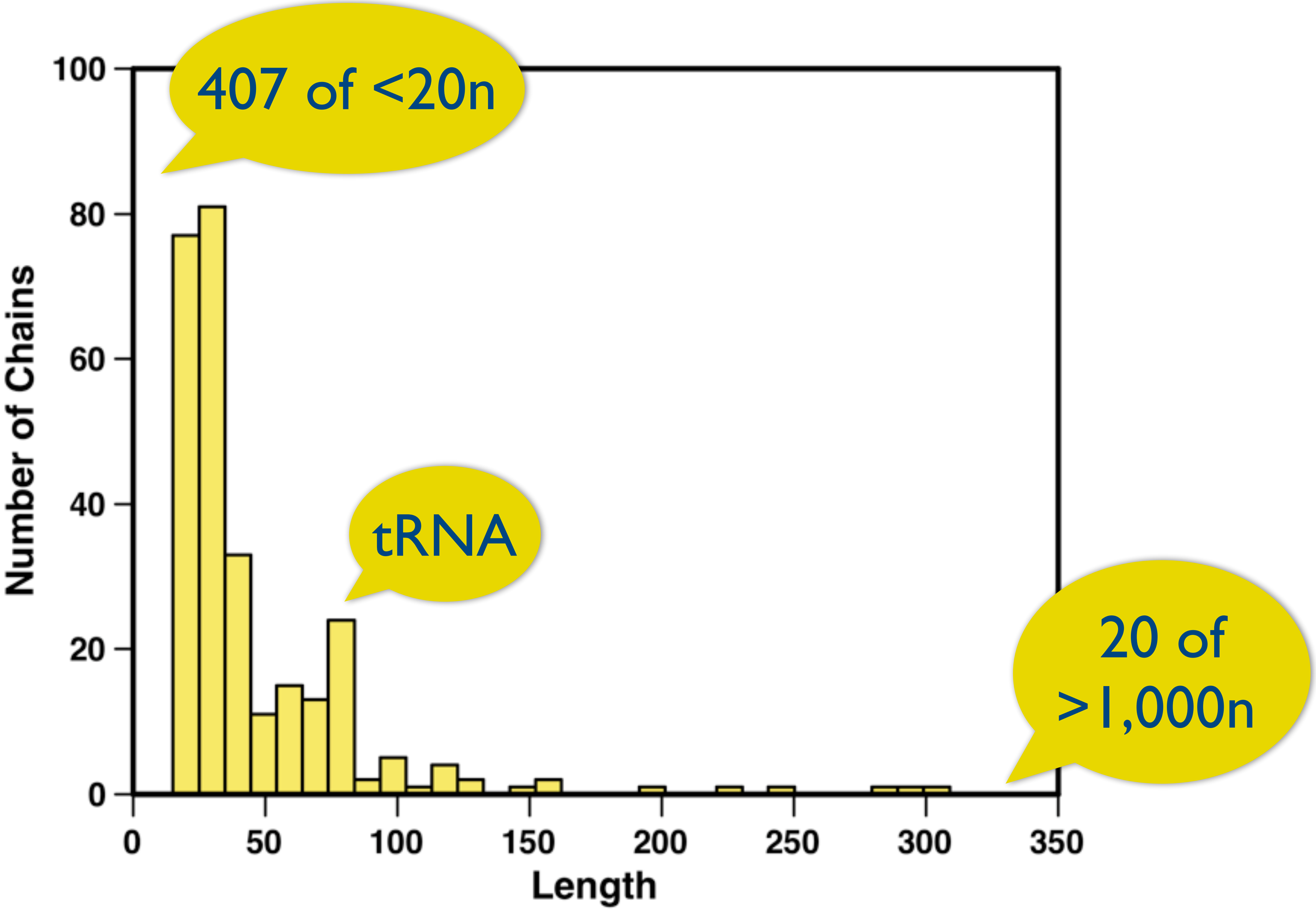
>Mutant Rat 28S rRNA sarcin/ricin domain
GGUGCUCAGUAUGAGAAGAACCGCACC
((((((((.((((..))))))))))))



Tertiary Structure

Secondary structure interactions and other interactions such as pseudoknots, hairpin-hairpin interactions, etc.

Dataset distribution



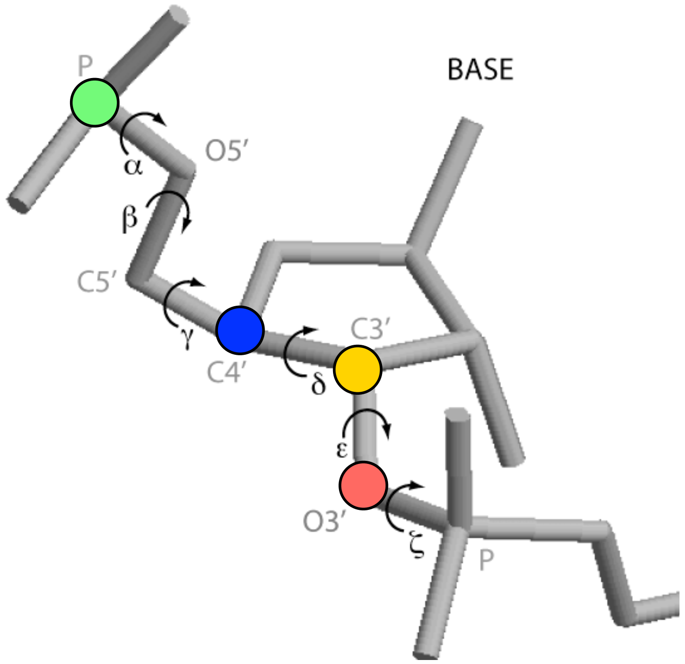
407 of <20n

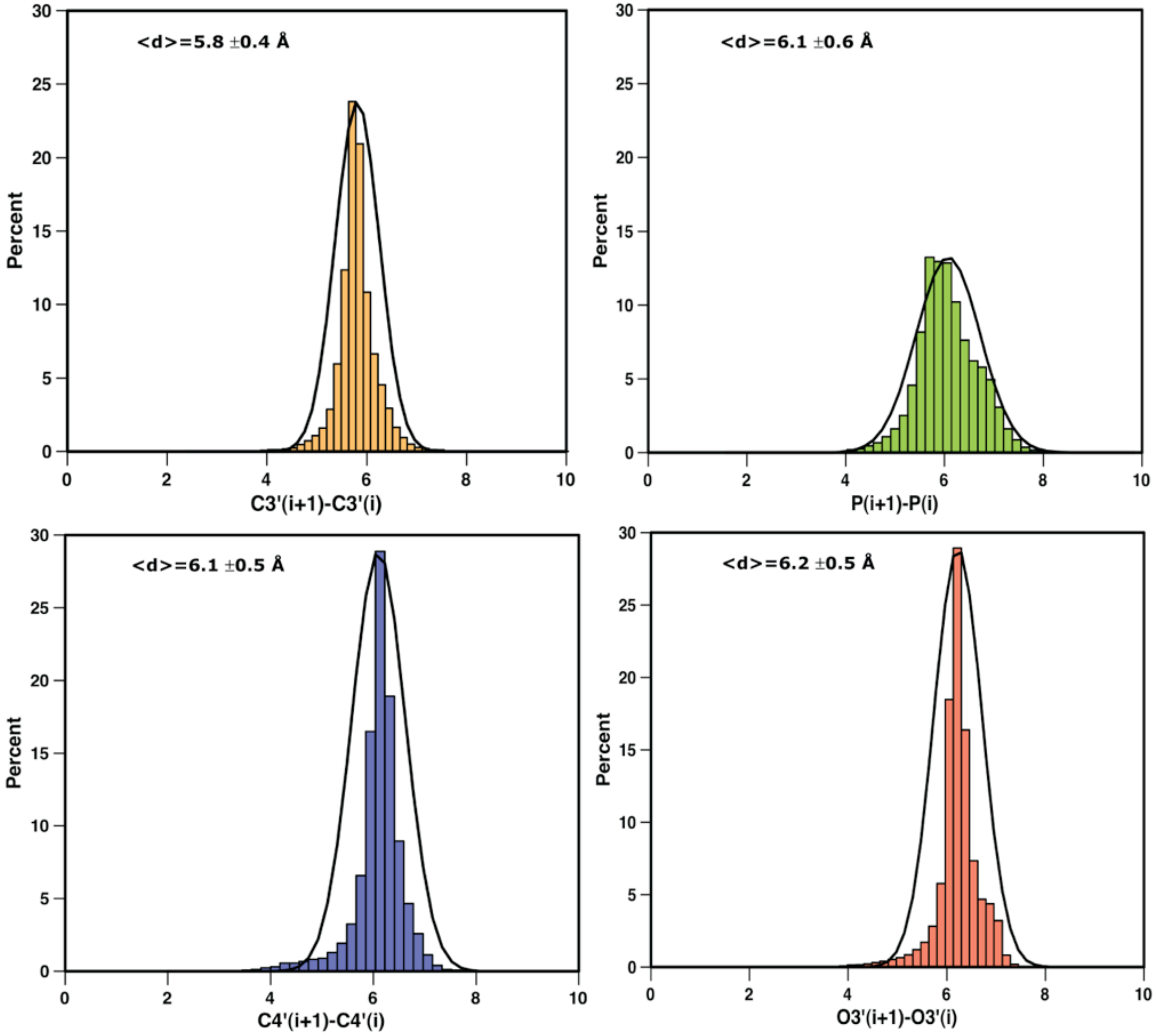
tRNA

20 of >1,000n

Atom selection

The best backbone atom that represents the RNA structure has been selected by evaluating the distribution of the distances between consecutive atoms in structures from the NR95 set.

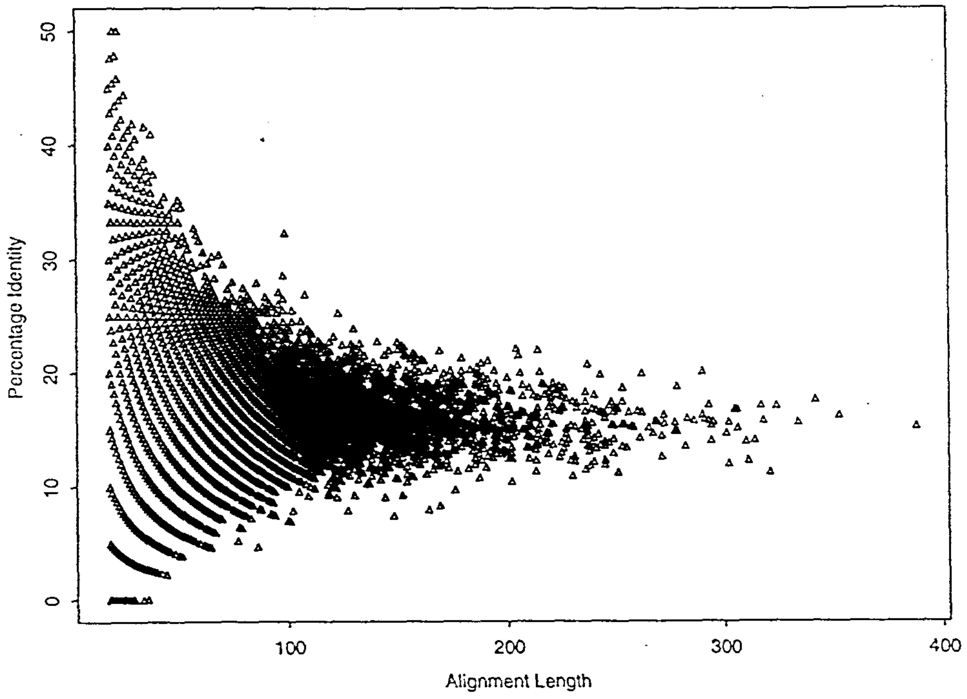




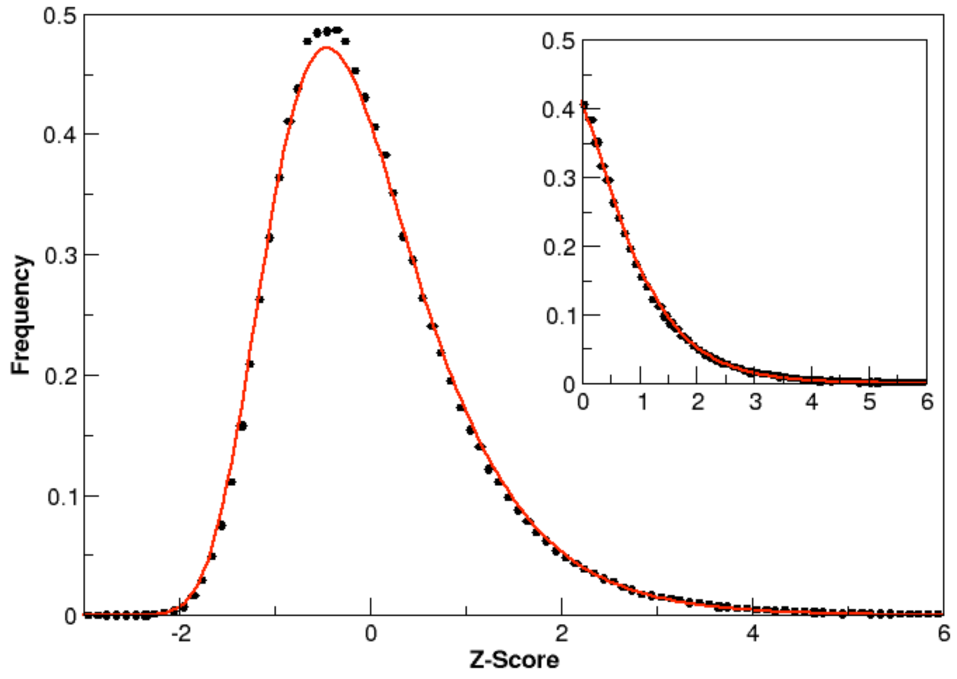
Background distribution

Considering a dataset of 300 random RNA structures, we have produced ~45,000 pairwise alignments that resulted in a empirical distribution. From such distribution we can then evaluate μ and σ needed to calculated the p-value for P(s>=x).

Empirical



Analytic



$$P(s \geq x) = 1 - \exp(-e^{-\lambda(s-\mu)})$$

Karlin and Altschul, (1990) PMID: 2315319

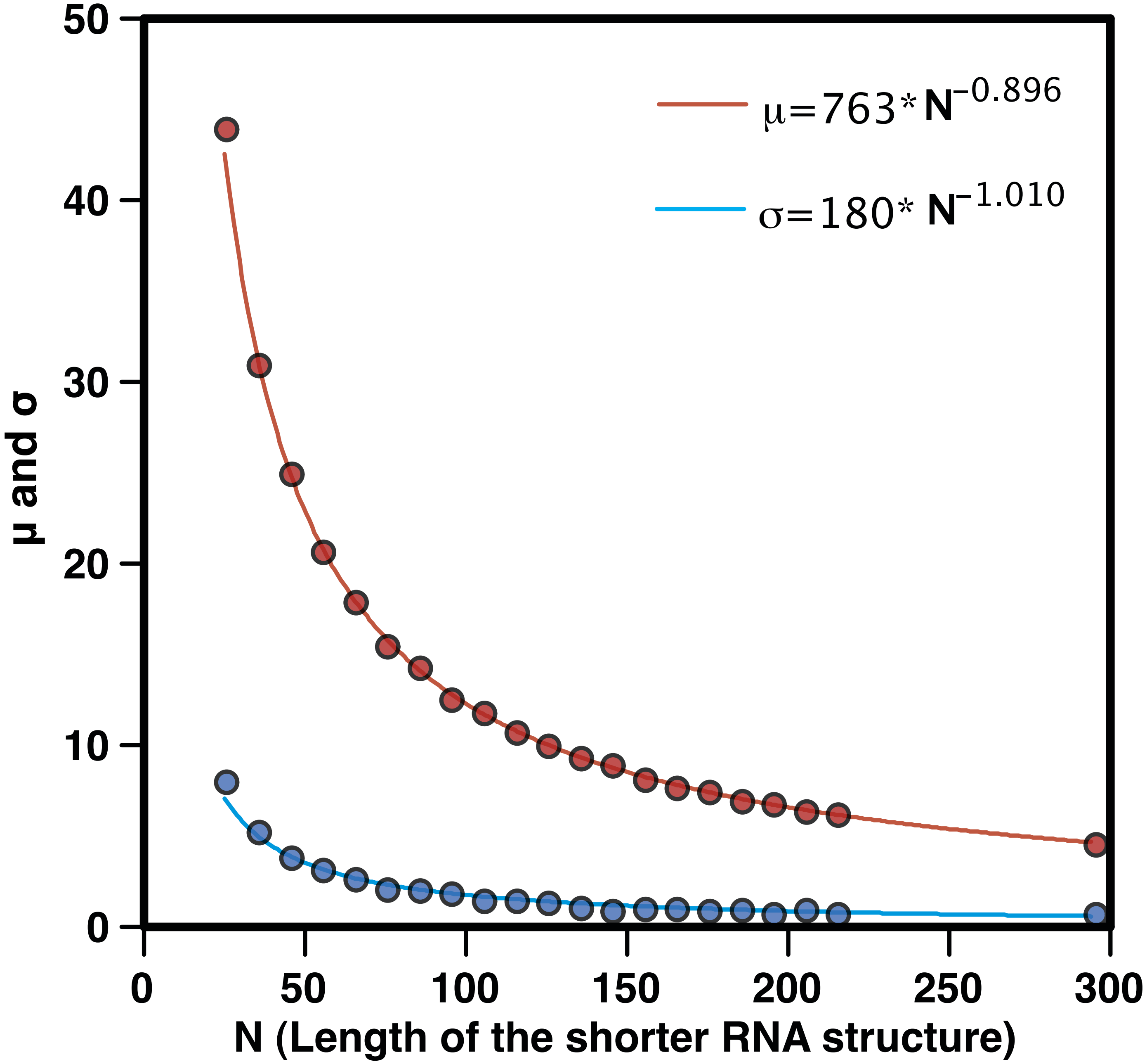
Mean and sigma

The score distribution depends on the length of the molecule.

We divided the resulting structural alignments (~45,000) in 30 bins according to the minimum sequence length of the two random structures (*N*).

For each bin the µ and σ values are evaluated fitting the data to an EVD.

The relations between *N* and µ, σ values are extrapolate fitting them to a power low function (r≈0.99).



Capriotti and Marti-Renom (2008) PMID: 18689811

Optimization

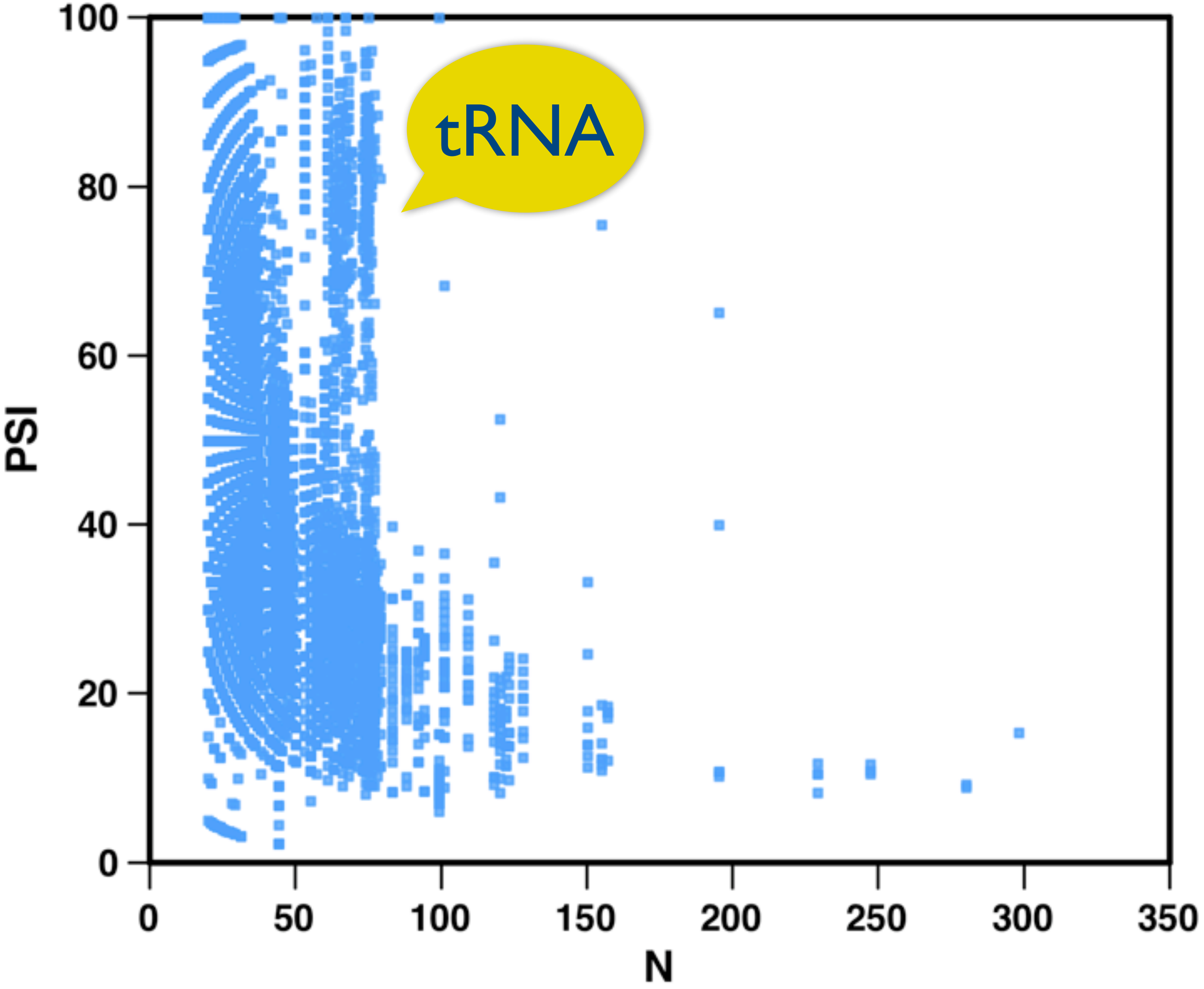
The accuracy of SARA method depends of a large number of parameters.

- C3’ and P backbone atoms for the unit vectors evaluation,
- k number of consecutive unit vectors, spamming from 3 to 9 and,
- values of gap opening from -9 to 0 and gap extension for -0.8 to 0
- Secondary structure information

	Gap opening	Gap extension	*k*
Secondary structure	-7.0	-0.6	3
No secondary structure	-8.0	-0.2	7

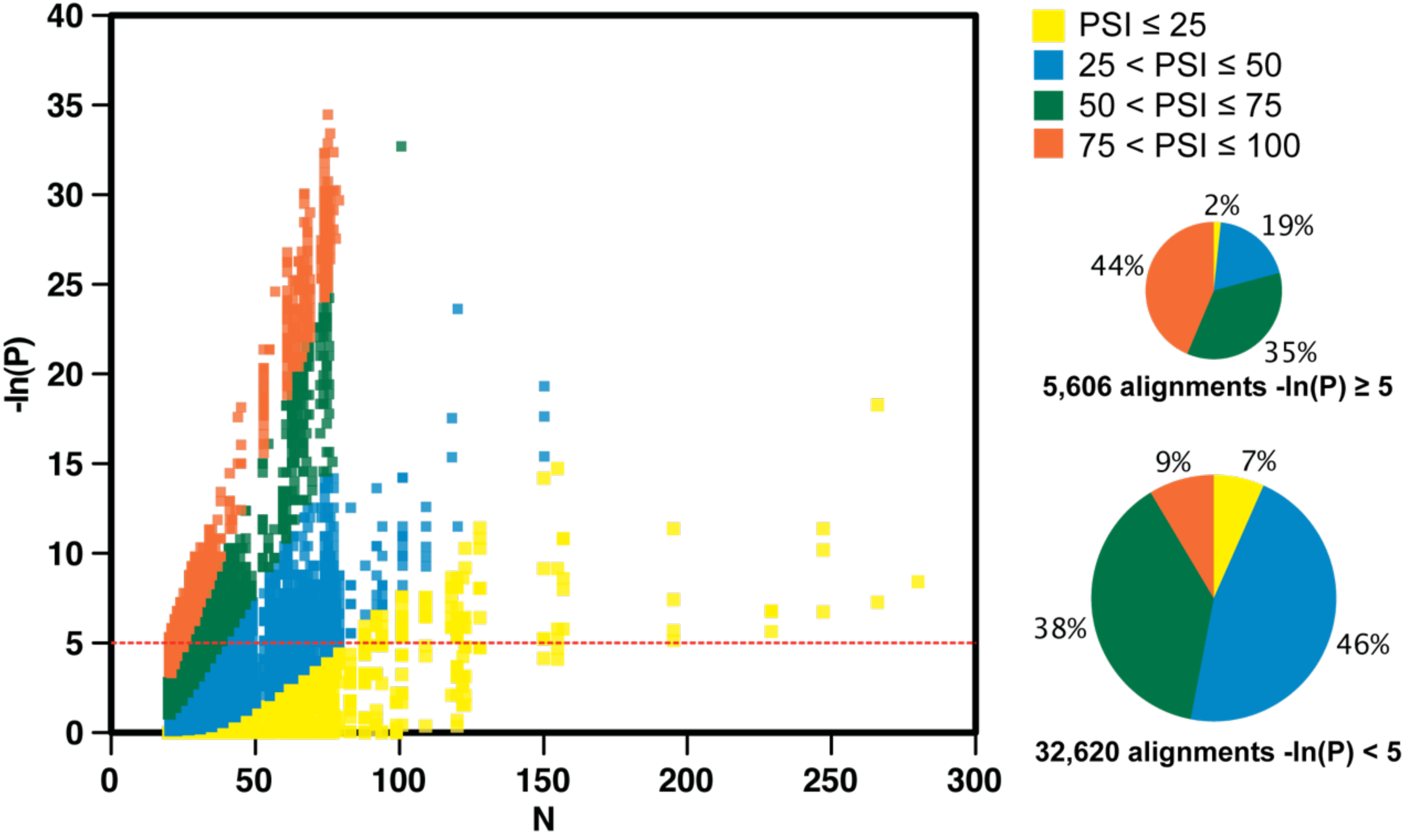
PSI distribution

all-against-all comparison of structures in the NR95 set



Statistical significance

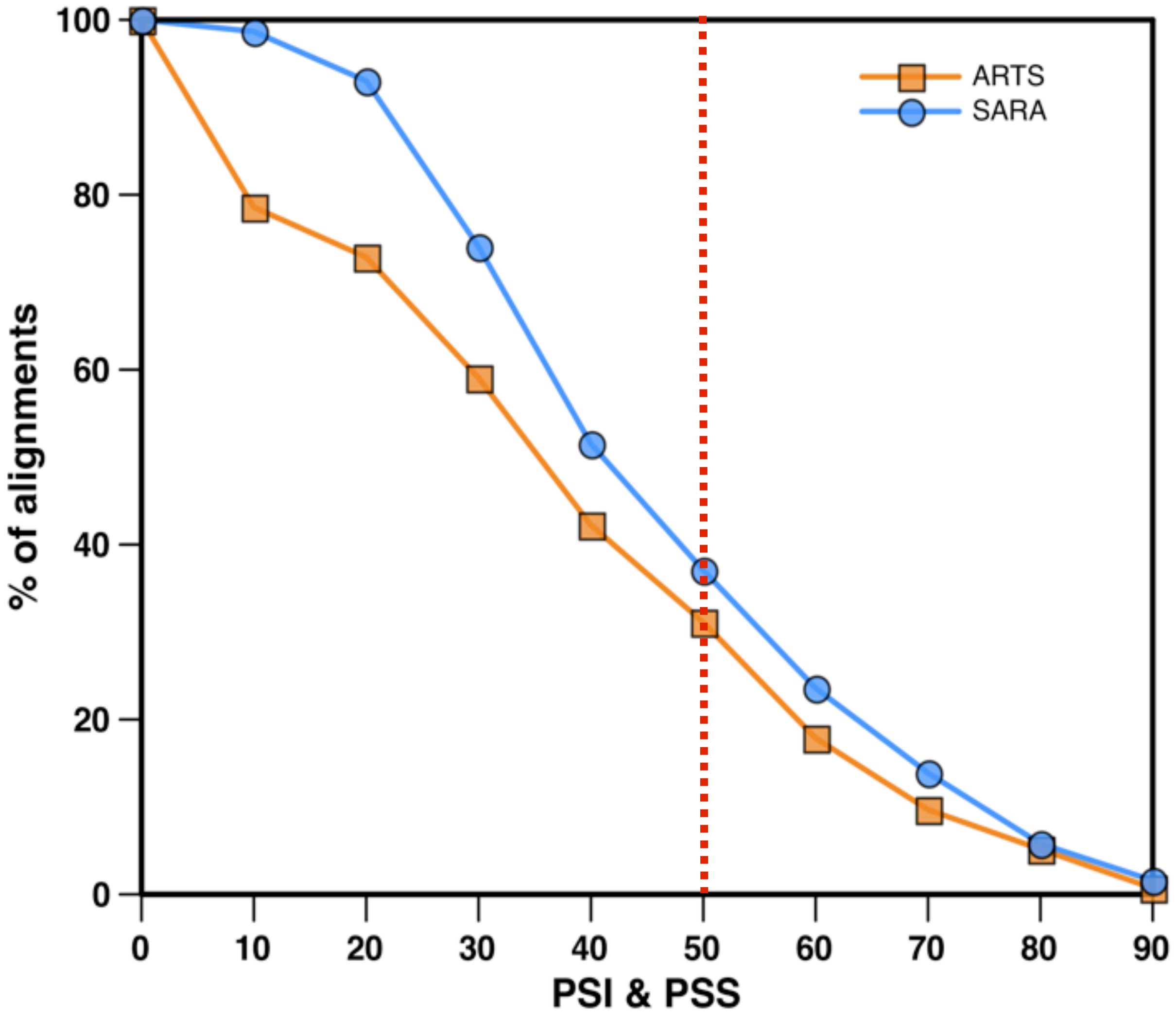
all-against-all comparison of structures in the NR95 set

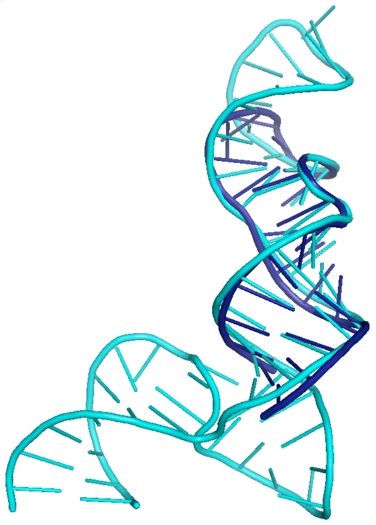


5,606 alignments -ln(P) ≥ 5

32,620 alignments -ln(P) < 5

Comparison with ARTS





SARA

Percentage of structure identity (PSI) 92.6%
Percentage of sequence identity 48.0%
Percentage of SSE identity 100.0%
RMSD 1.78 Å

```
>1q96 Chain:A
-------------------ggugcucaguaugag--------aagaaccgcacc-------
>1un6 Chain:E
gccggccacaccuacggggccugguuaguaccugggaaaccugggaauaccaggugccggc
```

ARTS

Percentage of structure identity (PSI) 76.9%
Percentage of sequence identity 20.0%
Percentage of SSE identity 79.2%
RMSD 1.66Å

```
>1q96 Chain:A
--------------------gugcucaguaugaga-----aga-accgcacc--------
>1un6 Chain:E
ccggccacaccuacggggccugguuaguaccugggaaaccugggaauaccaggugccggc
```

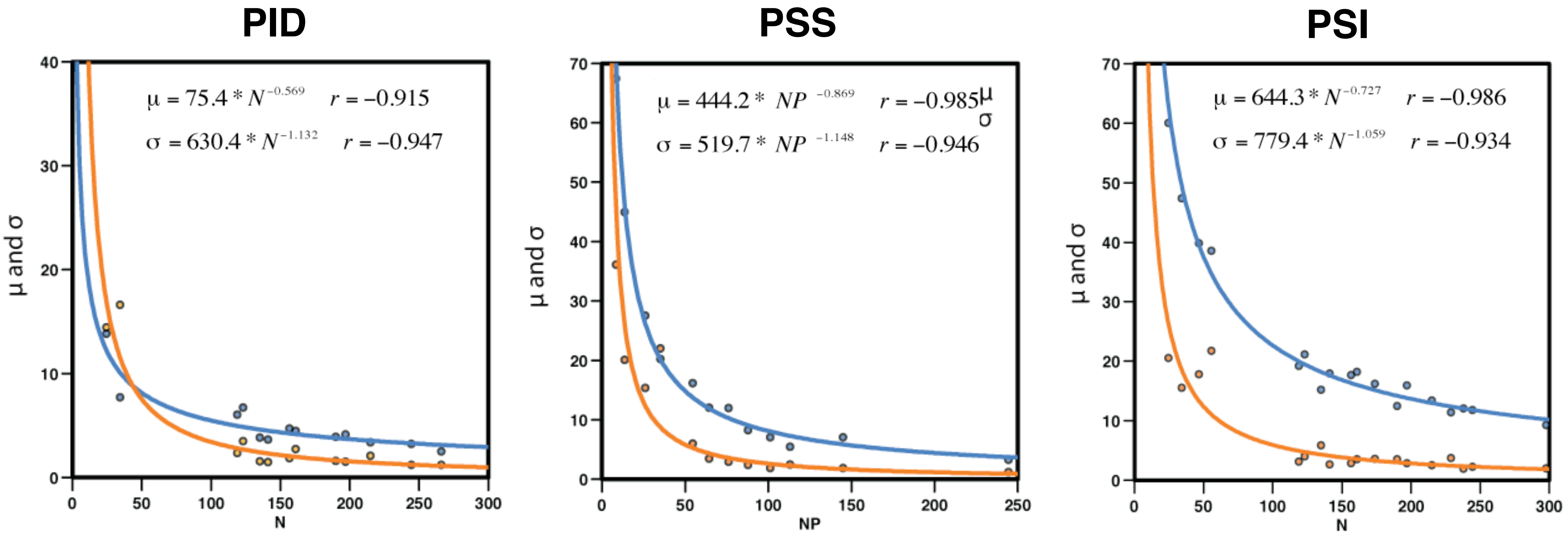
PSI: % of structure identity

PSS: % of secondary structure identity

Cut-off distance: 4.0 Å

Background distributions

Fitting of the μ and σ values. μ (blue) and σ (orange) parameters for PID, PSS and PSI that best fit an extreme value distribution. The distributions have been calculated using a set of 50,995 alignments between pairs of unrelated RNA.



Predicting RNA function

- The main idea behind this experiment is trying to predict RNA function using 3D structural alignments.

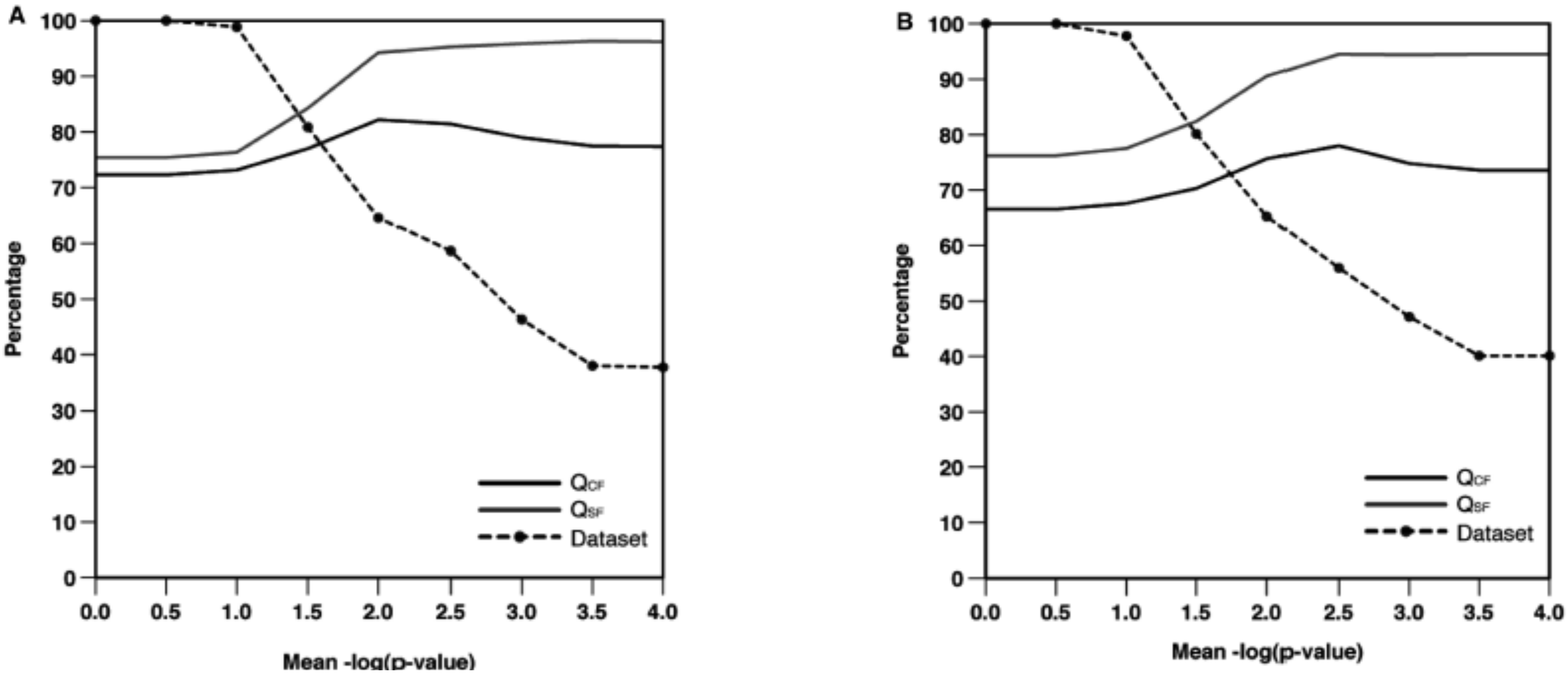
- We aligned an RNA structure with unknown function against the whole set of RNA structures annotated in SCOR database.

- The RNA function is inferred assigning the same function of the RNA the alignment with highest mean -ln(p-value).

- The method is tested using a leaving one out procedure on the whole annotated RNA structures in SCOR database.

Function assignment

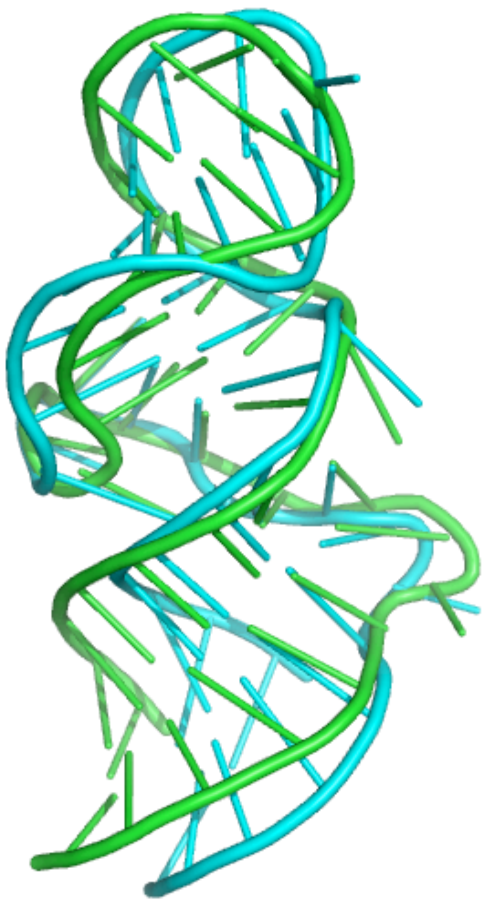
The accuracy of corrected function (Q_{CF}) and similar function (Q_{SF}) assignment tasks has been plotted as a function of the mean negative logarithm of the P-values for the best alignment. In (A) the plot results from leave one out on all SCOR set and (B) the performances using a representative SCOR subset

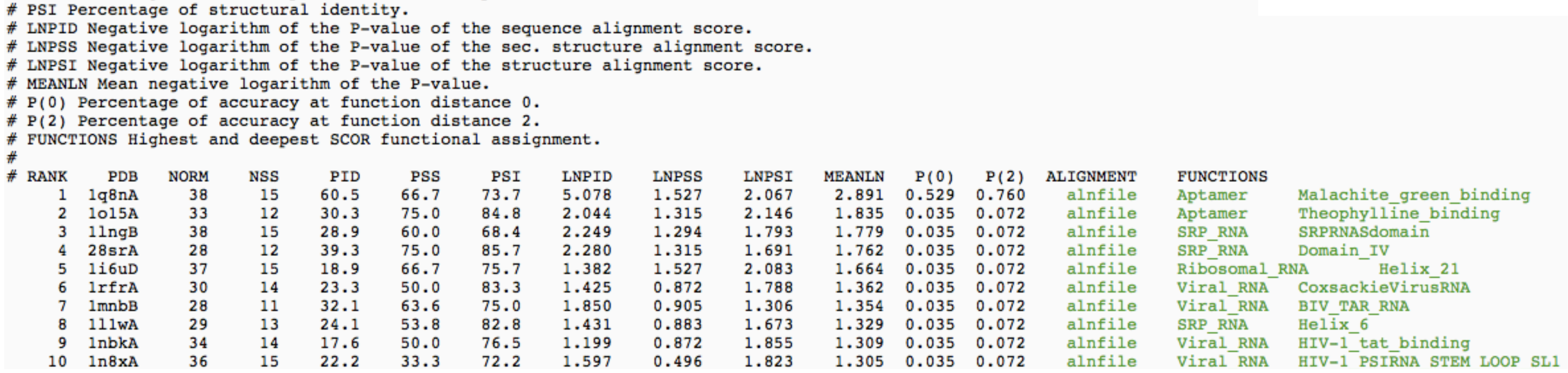


Capriotti and Marti-Renom. (2009), PMID: 19483098

Prediction example

1t1s chain A (cyan) is a RNA Aptamer that recognizes the chromophore malachite green. The structure ranked in the first position 1q8nA (green) has been classified as Malachite green binding Aptamer. The second structure is another Aptamer binding a different ligand.

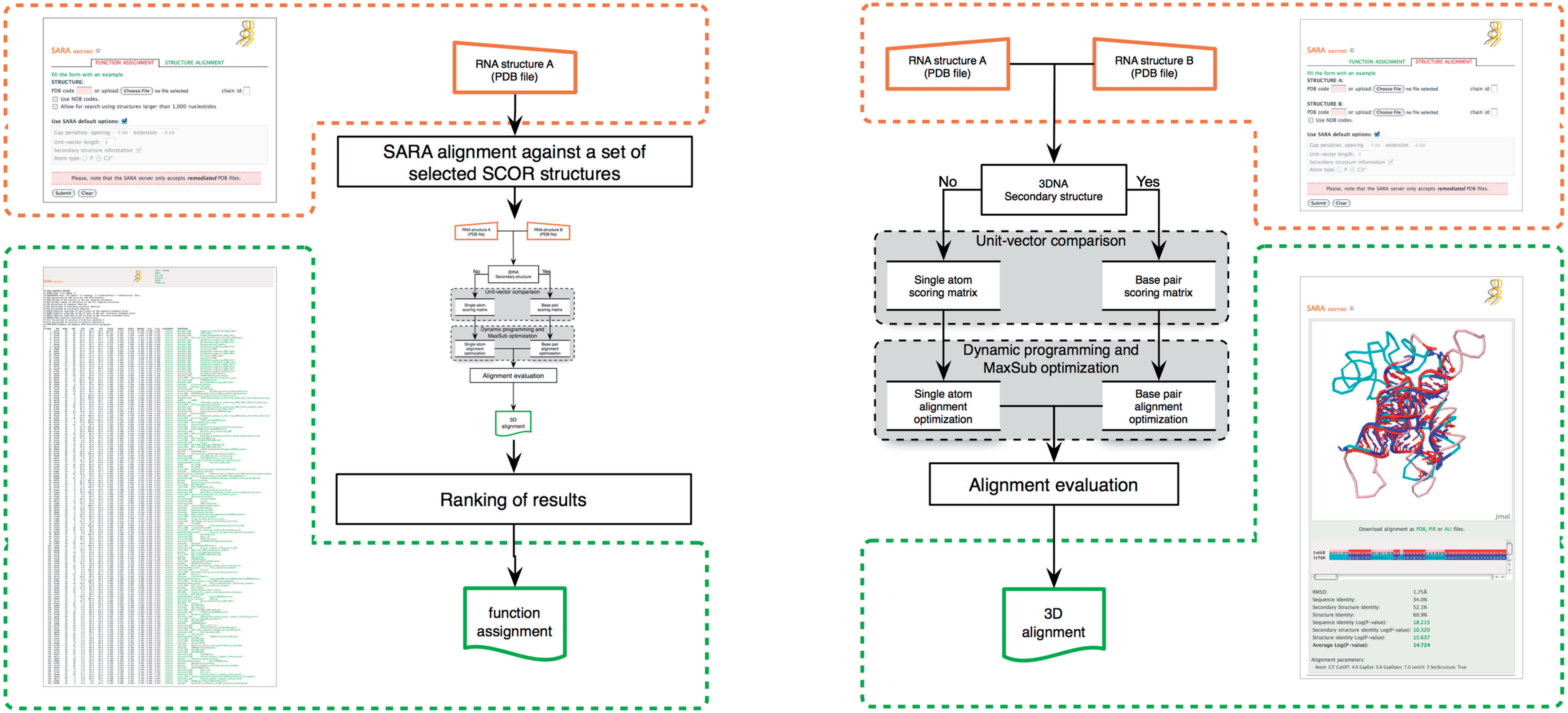




```
# SARA FUNCTION REPORT
# INPUT FILE: 1f1t CHAIN: A
# PARAMETERS Atom: C3' GapExt: 0.6 GapOpen: 7.0 LenUnitVector: 3 SecStructure: True
# PDB Representative PDB entry for the SCOR function.
# NORM Length of the shorter of the two compared structures.
# NSS Minimum number of base-pairs of the two compared structures.
# PID Percentage of sequence identity.
# PSS Percentage of secondary structure identity.
# PSI Percentage of structural identity.
# LNPID Negative logarithm of the P-value of the sequence alignment score.
# LNPSS Negative logarithm of the P-value of the sec. structure alignment score.
# LNPSI Negative logarithm of the P-value of the structure alignment score.
# MEANLN Mean negative logarithm of the P-value.
# P(0) Percentage of accuracy at function distance 0.
# P(2) Percentage of accuracy at function distance 2.
# FUNCTIONS Highest and deepest SCOR functional assignment.
#
# RANK    PDB   NORM    NSS     PID     PSS     PSI   LNPID   LNPSS   LNPSI  MEANLN  P(0)   P(2)  ALIGNMENT     FUNCTIONS
     1  1q8nA    38     15    60.5    66.7    73.7   5.078   1.527   2.067   2.891  0.529  0.760    alnfile     Aptamer     Malachite_green_binding
     2  1o15A    33     12    30.3    75.0    84.8   2.044   1.315   2.146   1.835  0.035  0.072    alnfile     Aptamer     Theophylline_binding
     3  1lngB    38     15    28.9    60.0    68.4   2.249   1.294   1.793   1.779  0.035  0.072    alnfile     SRP_RNA     SRPRNASdomain
     4  28srA    28     12    39.3    75.0    85.7   2.280   1.315   1.691   1.762  0.035  0.072    alnfile     SRP_RNA     Domain_IV
     5  1i6uD    37     15    18.9    66.7    75.7   1.382   1.527   2.083   1.664  0.035  0.072    alnfile     Ribosomal_RNA       Helix_21
     6  1rfrA    30     14    23.3    50.0    83.3   1.425   0.872   1.788   1.362  0.035  0.072    alnfile     Viral_RNA   CoxsackieVirusRNA
     7  1mnbB    28     11    32.1    63.6    75.0   1.850   0.905   1.306   1.354  0.035  0.072    alnfile     Viral_RNA   BIV_TAR_RNA
     8  1l1wA    29     13    24.1    53.8    82.8   1.431   0.883   1.673   1.329  0.035  0.072    alnfile     SRP_RNA     Helix_6
     9  1nbkA    34     14    17.6    50.0    76.5   1.199   0.872   1.855   1.309  0.035  0.072    alnfile     Viral_RNA   HIV-1_tat_binding
    10  1n8xA    36     15    22.2    33.3    72.2   1.597   0.496   1.823   1.305  0.035  0.072    alnfile     Viral RNA   HIV-1 PSIRNA STEM LOOP SL1
```

SARA server

The accuracy of corrected function (Q_{CF}) and similar function (Q_{SF}) assignment tasks has been plotted



RNA structure A (PDB file)
SARA alignment against a set of selected SCOR structures
Ranking of results
function assignment
RNA structure A (PDB file)
RNA structure B (PDB file)
No
3DNA Secondary structure
Yes
Unit-vector comparison
Single atom scoring matrix
Base pair scoring matrix
Dynamic programming and MaxSub optimization
Single atom alignment optimization
Base pair alignment optimization
Alignment evaluation
3D alignment
SARA server
FUNCTION ASSIGNMENT
STRUCTURE ALIGNMENT
Submit
Clear
Jmol

http://structure.biofold.org/sara

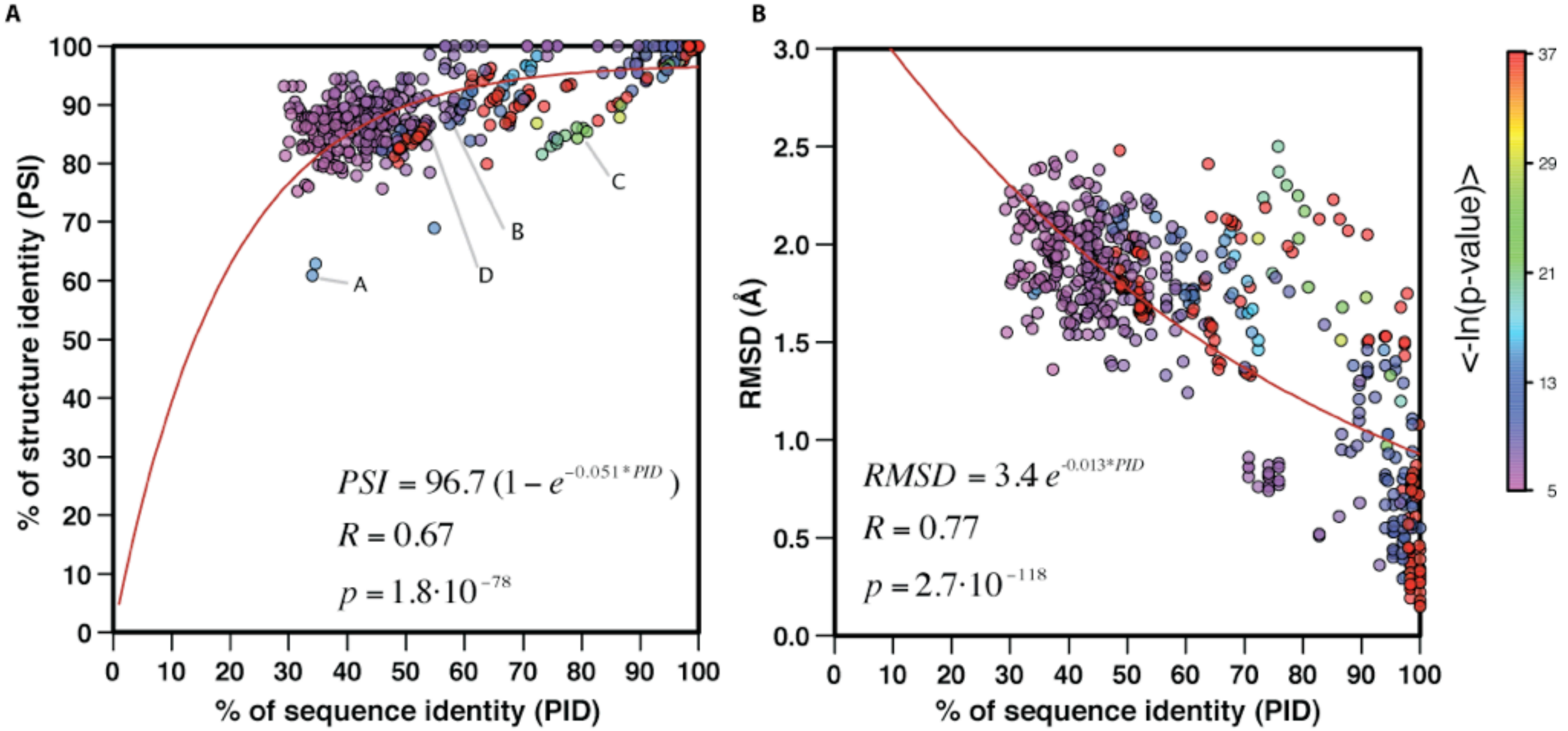
Capriotti and Marti-Renom. (2009), PMID: 19483098

Defining RNA structural space

- With the increasing number of available RNA structures we did the first attempt to define RNA structural space.
- We aligned aligned all against all a set of 451 non identical RNA structures and we selected a subset 589 high quality alignments.
- The relationship between sequence identity, secondary structure identity and 3D structure identity have be quantified
- We defined the twilight zone for RNA aligning all against all the sequences of same set of RNA using Infernal.

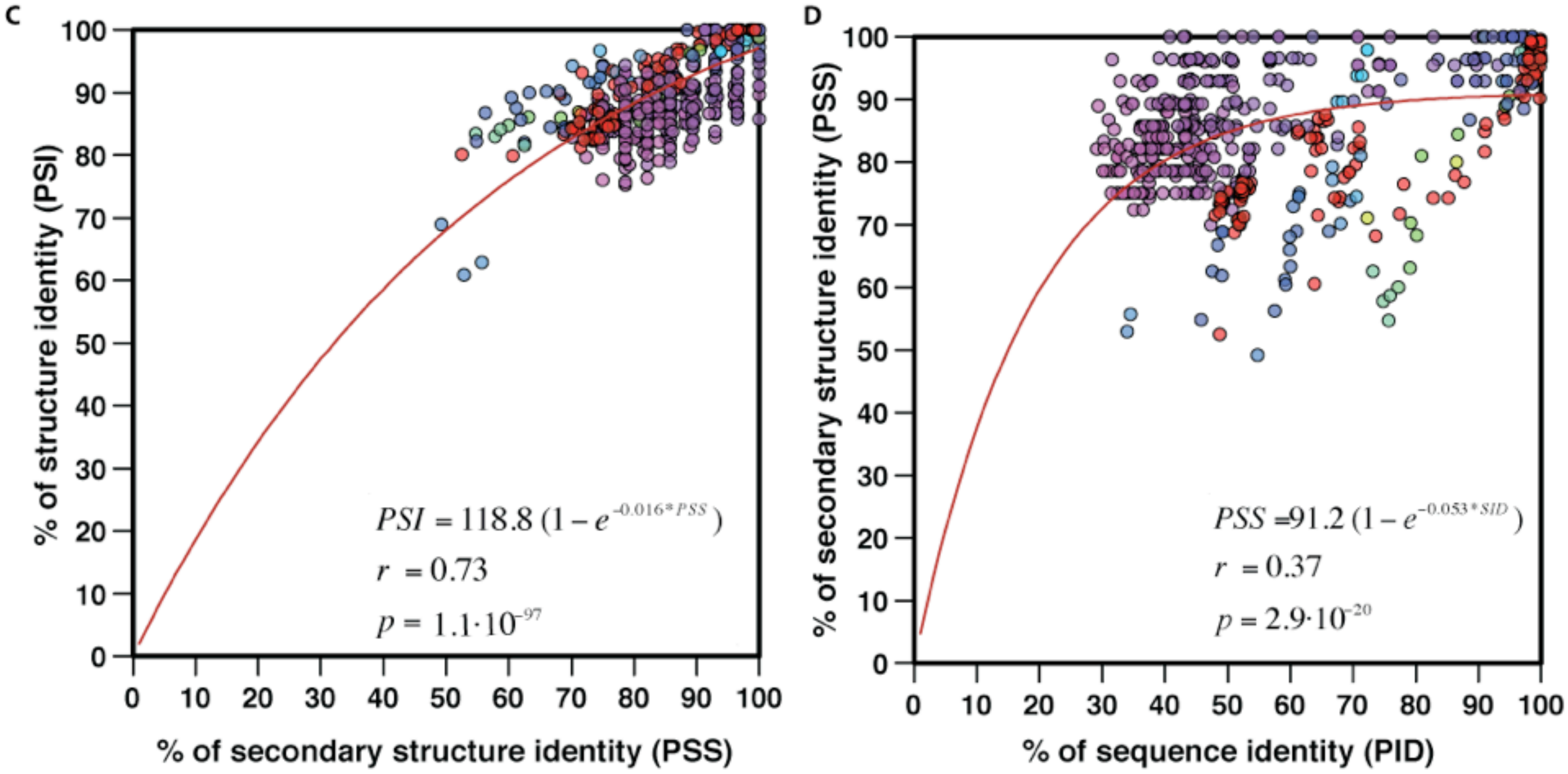
RNA structure space

The percentage of sequence identity (PID) correlates with the percentage of structure identity (PSI). Higher correlation coefficient is found between sequence identity and the RMSD value on the subset of atoms corresponding to equivalent residues. The correlation decreases in the region of sequence identity lower than 60%.



RNA secondary structure

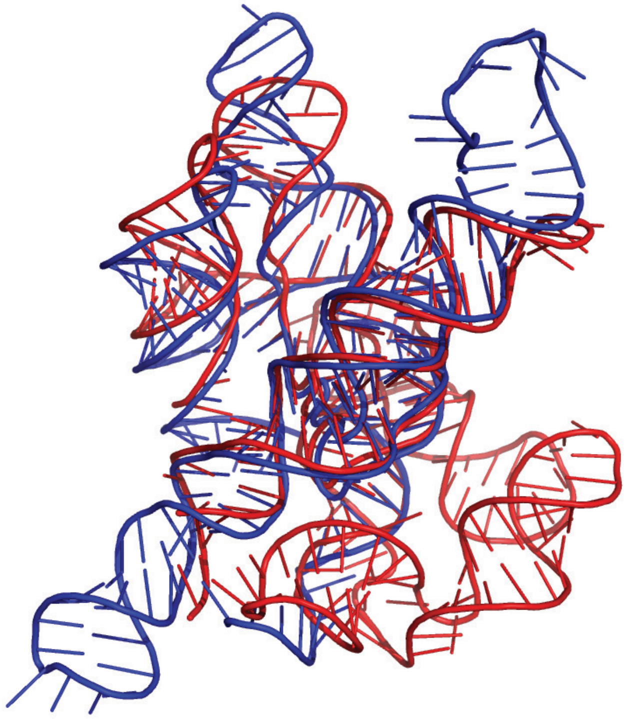
Secondary structure identity (PSS) strongly correlates with tertiary structure identity (PSI), meaning that good secondary structure alignments correspond to high tertiary structure similarity. The percentage of sequence identity (PID) poorly correlates with the percentage of secondary structure identity (PSS). This results is in agreement with low accuracy in the prediction of secondary structure.



Alignment examples (I)

Examples of medium quality RNA structural alignments for group I ribozyme and tRNA.

A Staphylococcus phage group I ribozyme (1y0q:A)
Synthetic I Intron fragment (1u6b:B)



Aligned nucleotides:	120
RMSD:	1.8 Å
Sequence Identity:	34.0 %
Secondary Structure Identity:	52.1 %
Structure Identity:	60.9 %
Sequence -ln(p-value):	18.2
Secondary structure -ln(p-value):	10.3
Structure -ln(p-value):	15.6
Mean -ln(p-value):	**14.7**

B Pyrococcus horikoshii tRNA(Leu) (1wz2:C)
Acuifex aeolicus tRNA(Met) (2ct8:C)

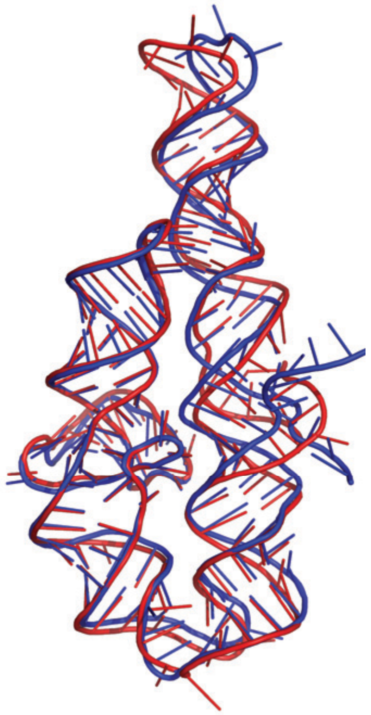
Aligned nucleotides:	65
RMSD:	1.9 Å
Sequence Identity:	56.8 %
Secondary Structure Identity:	88.5 %
Structure Identity:	87.8 %
Sequence -ln(p-value):	10.2
Secondary structure -ln(p-value):	5.2
Structure -ln(p-value):	7.2
Mean -ln(p-value):	**7.5**

Alignment examples (II)

Examples of high quality RNA structural alignments for P4-P6 RNA ribozyme and 23S RNA

C

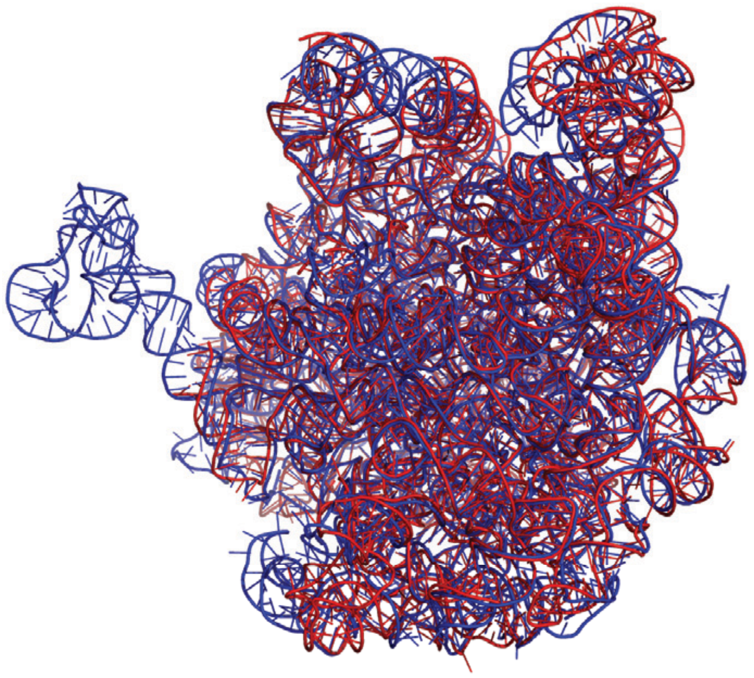
```
Synthetic P4-P6 RNA ribozyme (1l8v:A)
Synthetic P4-P6 RNA ribozyme (2r8s:R)
```



```
Aligned nucleotides:                   134
RMSD:                                  1.8 Å
Sequence Identity:                    80.9 %
Secondary Structure Identity:         81.0 %
Structure Identity:                   85.4 %
Sequence -ln(p-value):                37.0
Secondary structure -ln(p-value):     17.1
Structure -ln(p-value):               19.4
Mean -ln(p-value):                    24.5
```

D

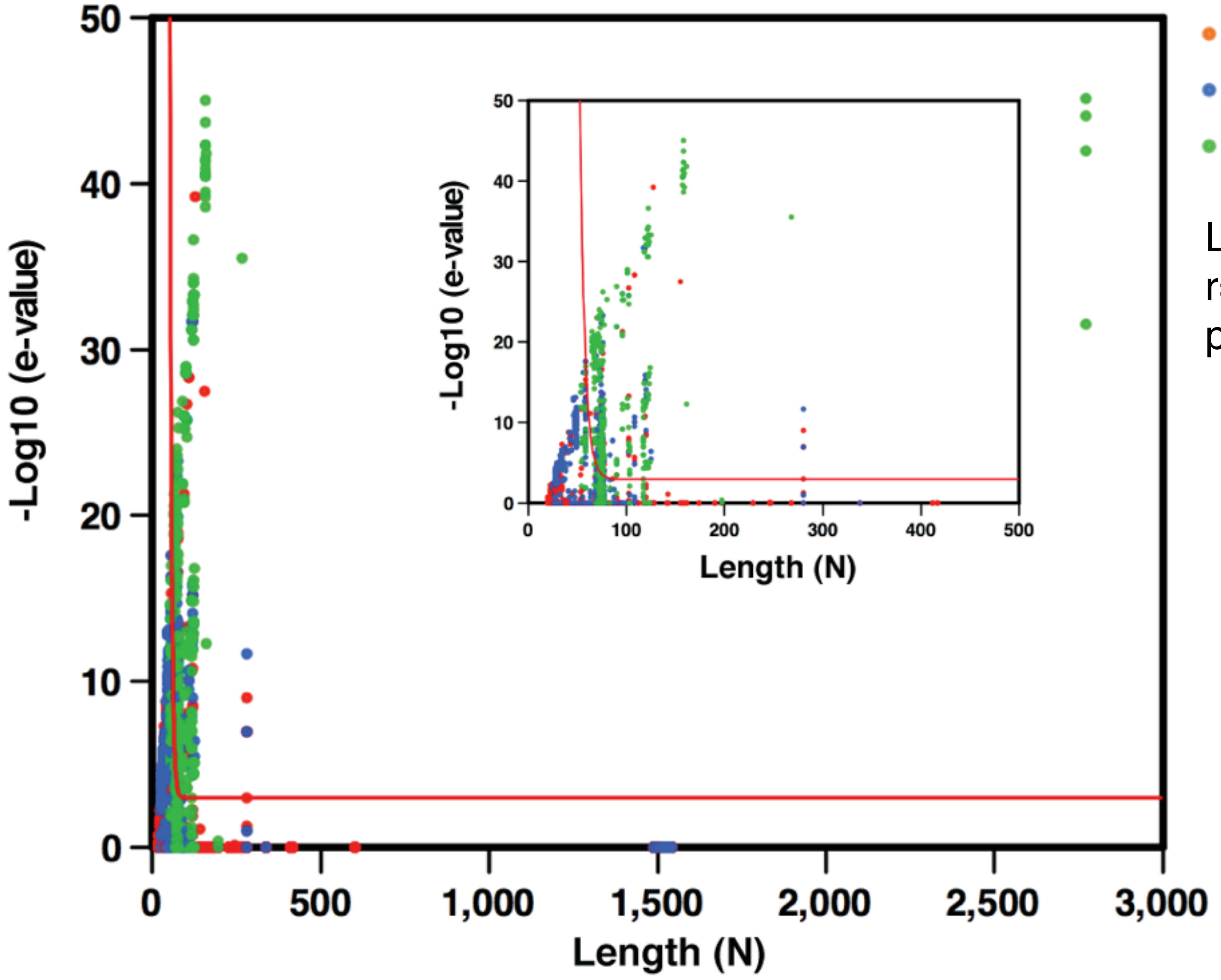
```
Haloarcula marismortui 23S RNA (3cce:0)
Thermus thermophilus 23S RNA (3d5b:A)
```



```
Aligned nucleotides:                 2,347
RMSD:                                  1.7 Å
Sequence Identity:                    52.7 %
Secondary Structure Identity:         75.7 %
Structure Identity:                   85.2 %
Sequence -ln(p-value):                37.0
Secondary structure -ln(p-value):     37.0
Structure -ln(p-value):               37.0
Mean -ln(p-value):                    37.0
```

RNA twilight zone

It is possible to calculate the twilight-zone curve that better discriminates between high and low quality alignments.



● All -log(p-values) ≤ 4.5

● At least one -log(p-value) ≤ 4.5

● All -log(p-values) > 4.5

$\mathrm{Log}_{10}(\text{e-value})=3.11+362*e^{-0.08N}$

$r=0.85$

$p=1.6*10^{-2}$

Capriotti and Marti-Renom (2010) PMID: 20550657