The interval branch-and-prune algorithm for the protein structure determination by NMR

Thérèse E Malliavin

Unité de Bioinformatique Structurale
Institut Pasteur and UMR CNRS 3528
Paris, France

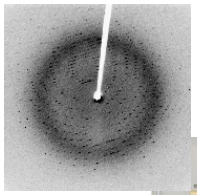
DIMACS Workshop on Distance Geometry: Theory and Applications
Rutgers University, 26-29 July 2016

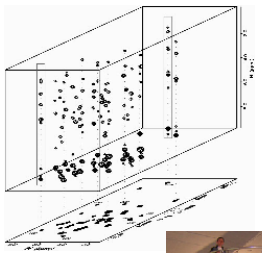
Experimental techniques for structural biology

MOLECULAR SIZE

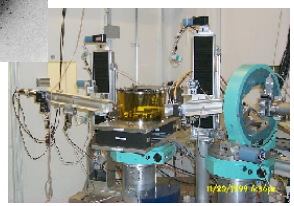
DISORDER

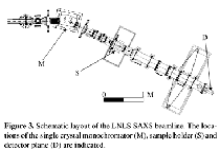
Nuclear Magnetic Resonance



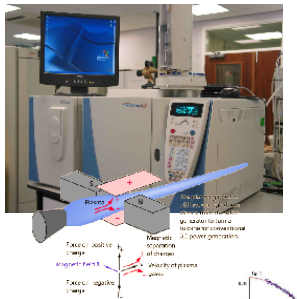


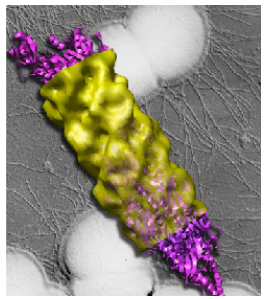
X-ray crystallography

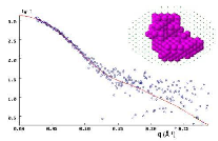




Cryo-EM Microscopy







Small-angle X-ray scattering

The infancy of structural biology

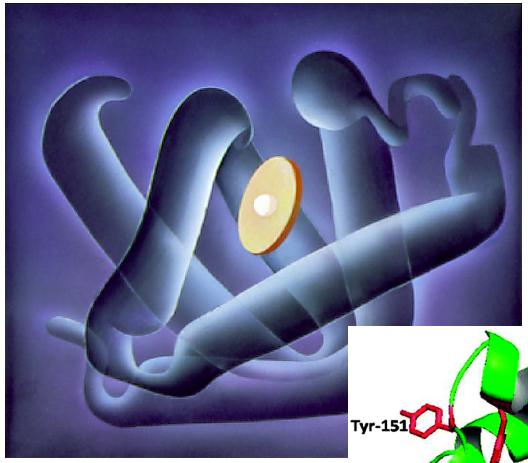
Model of pig insulin: Dorothy Hodgkin, 1967 (Image credit: Science Museum).

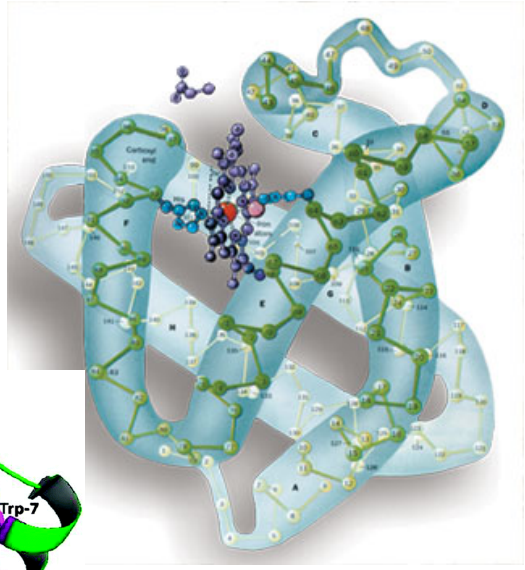
John Kendrew with the 'forest of rods' model of myoglobin (Image credit: MRC Laboratory of Molecular Biology)

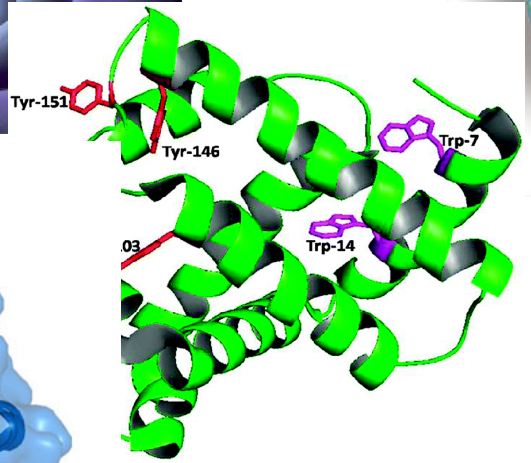
Sausage model of myoglobin, John Kendrew, 1957 (Image credit: Science Museum).

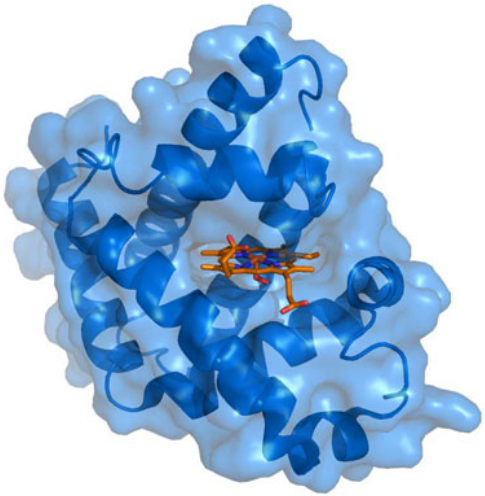
kathryngamer.co.uk/blog

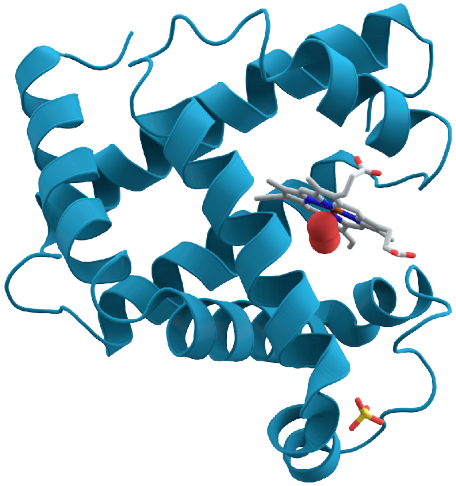
Structural bioinformatics









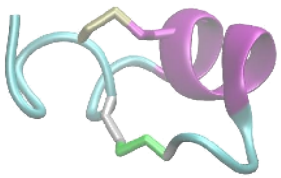


Low ordered or disordered biomolecular structure

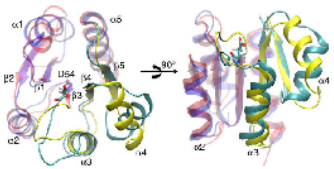
« Unique » conformation

Conformational exchange

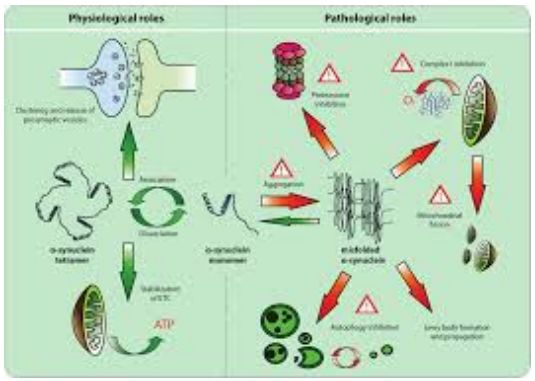
Instrinsically Disordered Proteins (IDP)



Conotoxin, 1IEN
Sharp...Lewis
Nat Neurosc 2001



nitrogen regulatory protein C (NtrC). Vanatta...Pande, Nat Comm 2014



Bobela... Schneider, Biomolecules 2015

X-ray crystallography

Nuclear Magnetic Resonance (solid-state, solution)

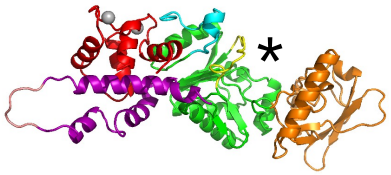
Small-angle X-ray scattering

Fluorescence resonance energy transfer

Exploration of AC conformational space

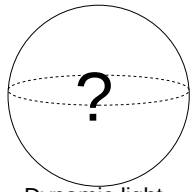
- AC inactive state
- Larger range of explored gyration radii

Bordetalla pertussis



Active enzyme

Inactive enzyme

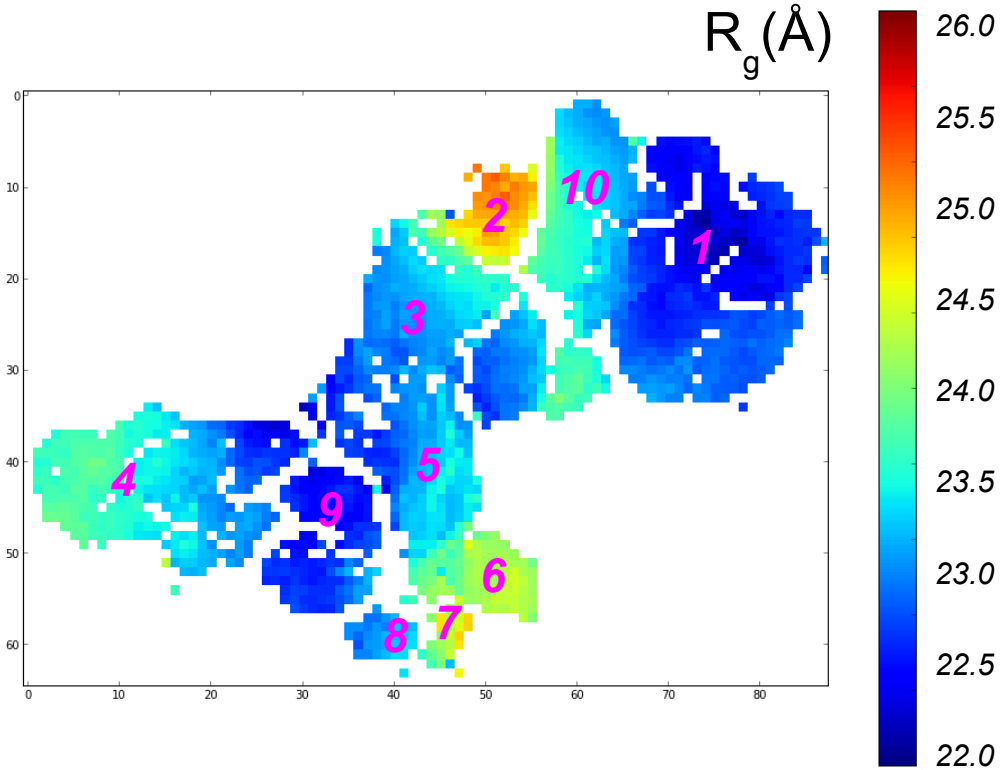


Dynamic light scattering
Karst et al, Biochem 2010

Cortes-Ciriano, Bouvier, Nilges, Maragliano, Malliavin. Temperature Accelerated Molecular Dynamics with Soft-Ratcheting Criterion Orients Enhanced Sampling by Low-Resolution Information. J Chem Theory Comput 2015.

Exploration of AC conformational space

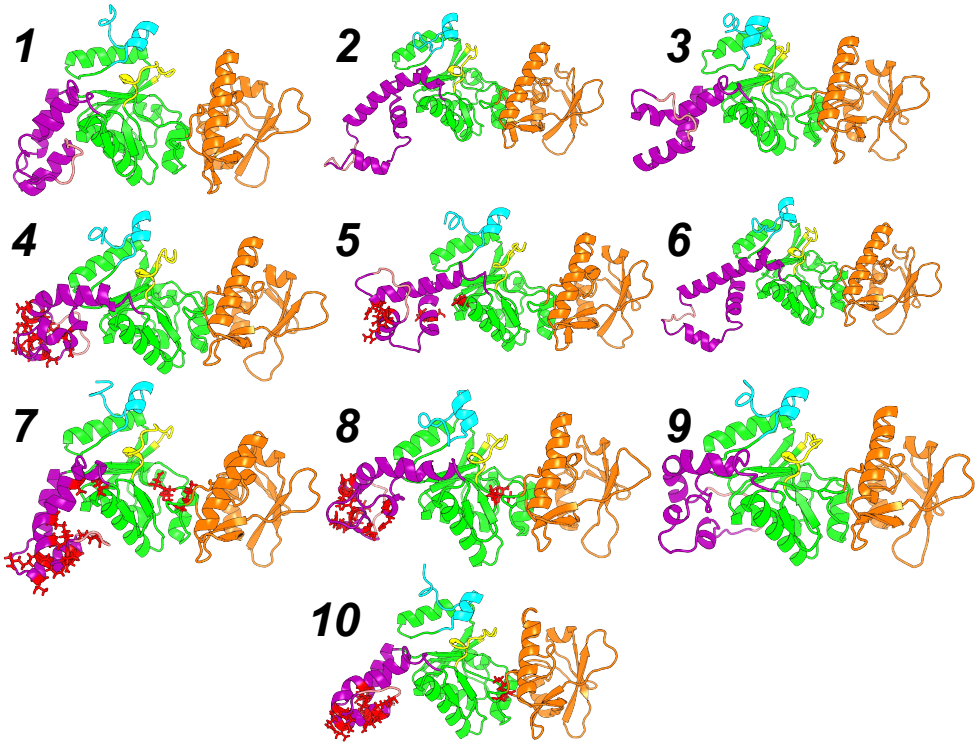
- AC inactive state
- Larger range of explored gyration radii
- Clustered compact conformations



Cortes-Ciriano, Bouvier, Nilges, Maragliano, Malliavin. Temperature Accelerated Molecular Dynamics with Soft-Ratcheting Criterion Orients Enhanced Sampling by Low-Resolution Information. J Chem Theory Comput 2015.

Exploration of AC conformational space

- AC inactive state
- Larger range of explored gyration radii
- Clustered compact conformations



Cortes-Ciriano, Bouvier, Nilges, Maragliano, Malliavin. Temperature Accelerated Molecular Dynamics with Soft-Ratcheting Criterion Orients Enhanced Sampling by Low-Resolution Information. J Chem Theory Comput 2015.

Global optimisation

Global optimization is distinguished from regular optimization by its focus on finding the maximum or minimum over all input values, as opposed to finding local minima or maxima

Martin,

Zhou, Donald. Systematic solution to homo-ligomeric structures determined by NMR Proteins 2015.

Distance Geometry problem

- Problem
- Spheres intersection

At the beginning of this discussion, we will suppose that all distances in G are precise. In this case, the MDGP can be seen as the problem of finding a conformation $x = (x_1, x_2, \ldots, x_n)$ such that all constraints:

$$||x_u - x_v|| = d(u, v) \qquad \forall (u, v) \in E \tag{16.1}$$

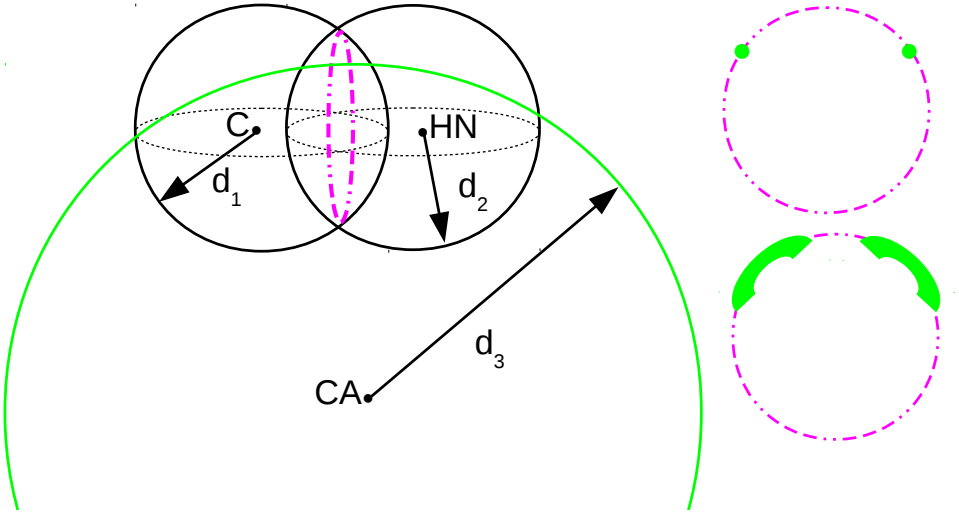
are satisfied. In the formula, $||\cdot||$ represents the computed distance between two atomic coordinates belonging to the conformation x, whereas $d(u, v)$ represents the known distance between the two atoms (the weight associated to the edge). The MDGP is a constraint satisfaction problem.

Malliavin, Mucherino, Nilges. Distance geometry in structural biology: new perspectives in: Distance Geometry: Theory, Methods and Applications, Mucherino, Lavor, Liberti, Maculan (Eds.), Springer 2013.

Distance Geometry problem

- Problem
- Spheres intersection
- Tree: branch and prune

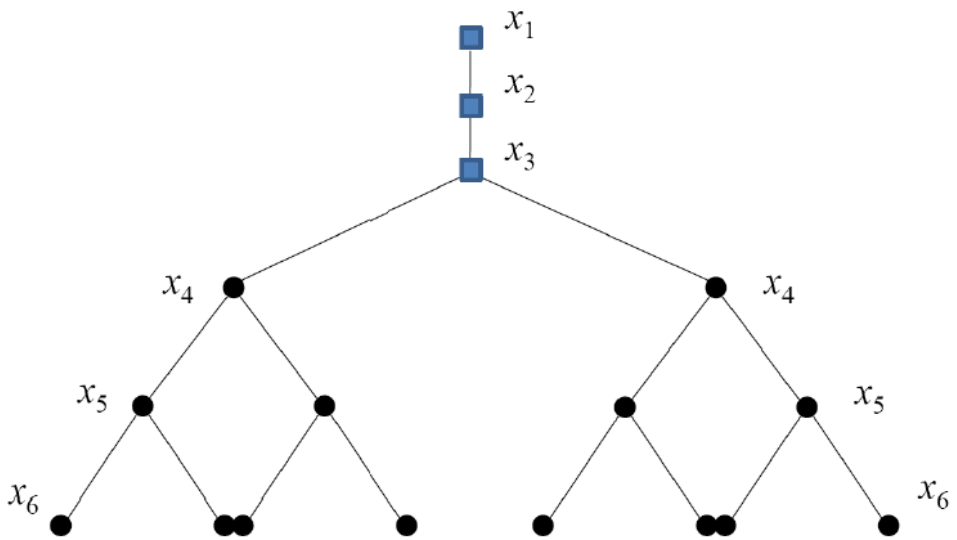
Search for the position of an atom X such that:
dist(X,C) = d_1, dist(X,HN) = d_2, dist(X,CA) = d_3



Malliavin, Mucherino, Nilges. Distance geometry in structural biology: new perspectives in: Distance Geometry: Theory, Methods and Applications, Mucherino, Lavor, Liberti, Maculan (Eds.), Springer 2013.

Distance Geometry problem

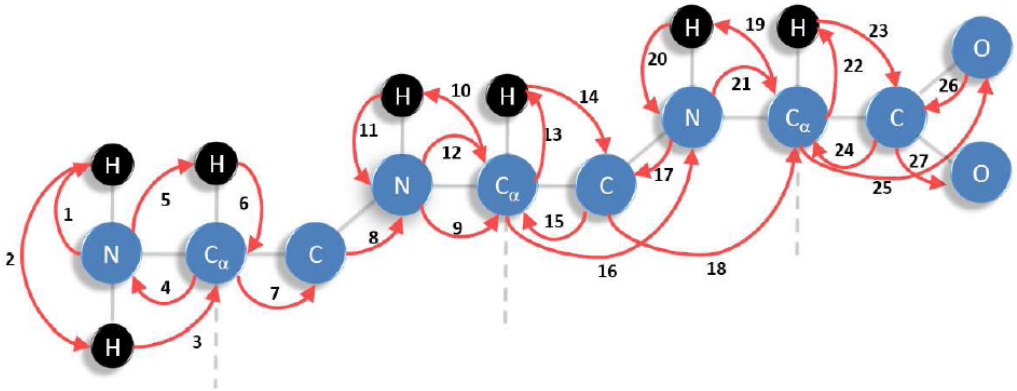
- Problem
- Spheres intersection
- Tree: branch and prune
- Atoms recursive ordering



Malliavin, Mucherino, Nilges. Distance geometry in structural biology: new perspectives in: Distance Geometry: Theory, Methods and Applications, Mucherino, Lavor, Liberti, Maculan (Eds.), Springer 2013.

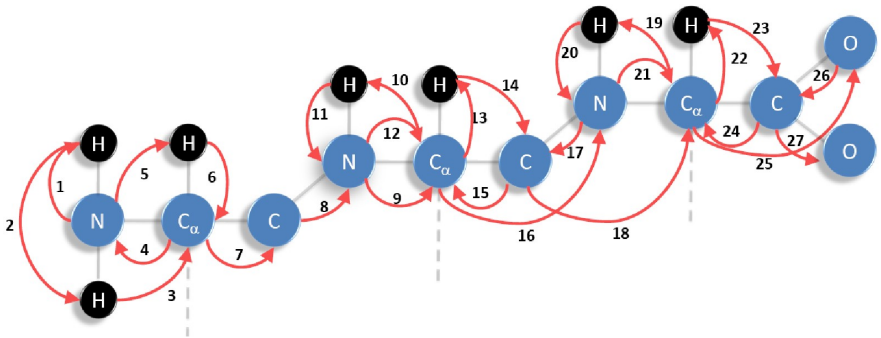
Distance Geometry problem

- Problem
- Spheres intersection
- Tree: branch and prune
- Atoms recursive ordering



Malliavin, Mucherino, Nilges. Distance geometry in structural biology: new perspectives in: Distance Geometry: Theory, Methods and Applications, Mucherino, Lavor, Liberti, Maculan (Eds.), Springer 2013.

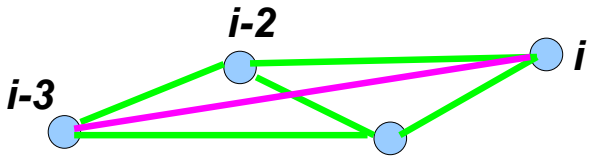
Atom recursive ordering and branching distances



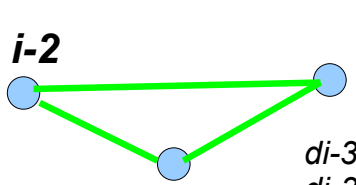
Distances corresponding to chemical bond, or to 2 chemical bonds connected by a bond angle

Distances that can be exact or interval

Case without cycle

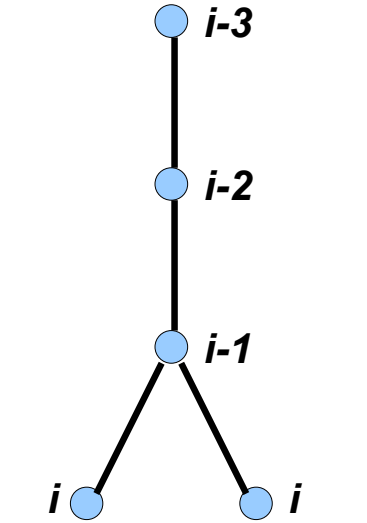


Case with cycle



di-3,i = 0
di-3,i-2 = di,i-2
cos(ω) = 1
sin(ω) = 0

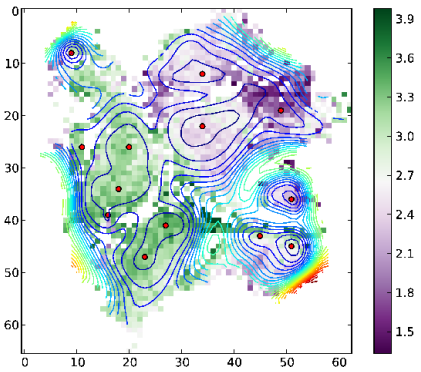
Atom ordering and branching



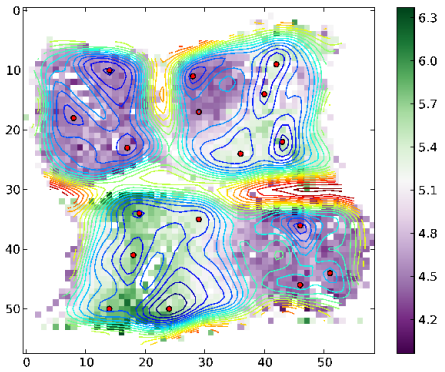
Helical peptides with long-range distance restraints

- Exploration of conformations using few long-range distance restraints
- Superposition to Protein Data Bank target structures

2KXA (aa)



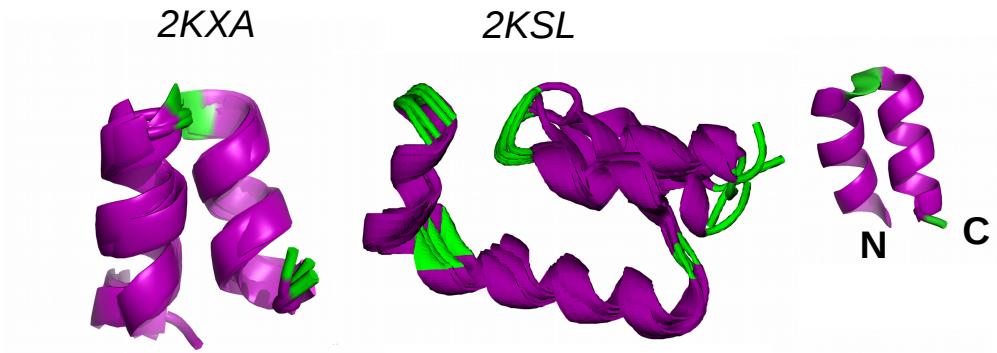
2KSL (aaaa)



Cassioli, Bardiaux, Bouvier, Mucherino, Alves, Liberti, Nilges, Lavor C, Malliavin TE. BMC Bioinformatics 2015

Helical peptides with long-range distance restraints

- Exploration of conformations using few long-range distance restraints
- Superposition to Protein Data Bank target structures



PROCHECK	2KXA	2KSL
Core	63.6	75.5
Allowed	36.4	14.2
Generously Allowed	0.0	9.2
Disallowed	0.0	1.0

Cassioli, Bardiaux, Bouvier, Mucherino, Alves, Liberti, Nilges, Lavor C, Malliavin TE. BMC Bioinformatics 2015

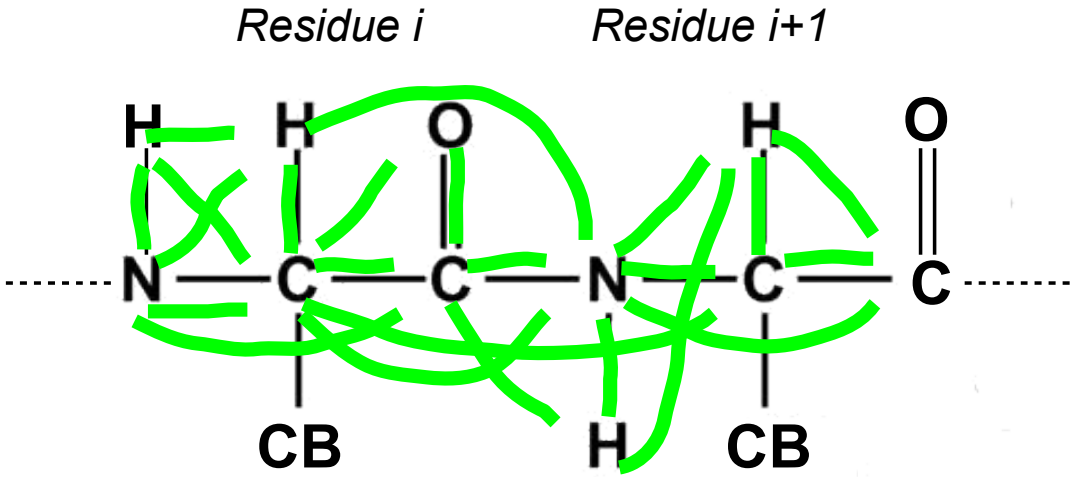
Validation set of proteins

PDB	Nres	SecStruct	EXP
1CEY	128	abababababab	NMR
2F05	85	aaaa	NMR
2KSL	51	aaaa	NMR
2KXA	24	aa	NMR
2LJ0	65	bbbbb	NMR
2LVR	30	bba	NMR
2LXZ	32	bbb	NMR
2M5X	40	bbab	NMR
2MC6	73	bbba	NMR
2MDI	56	bbb	NMR
2MGV	65	abbb	NMR
2MH2	64	abaabb	NMR

PDB	Nres	SecStruct	EXP
2MJ6	90	aabbbb	NMR
2MLA	37	babb	NMR
2MNI	92	baabba	NMR
2MP1	77	bbba	NMR
2MW9	33	bbb	NMR
2MXE	47	bbaab	NMR
2N17	56	bbab	NMR
2N2Q	54	babb	NMR
2RUP	58	bbb	NMR
4BYA	75	aaaa	NMR
4OU0	66	abaabb	XR
4RBX	32	bbb	XR

Calculations using exact distances and intervals, determined from a limited informative set of restraints.

Short-range exact distance are used



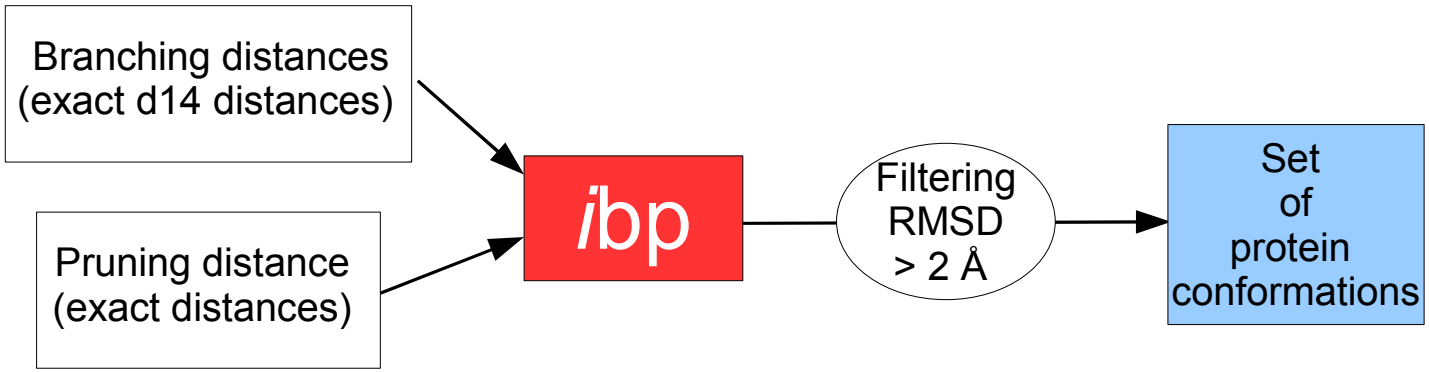
Exact branching distances

Exact pruning distances: involved residues

p = 5,4: (i,i) (i,i+1)
p = 6,4: (i,i) (i,i+1)
p = 7,4: (i,i) (i,i+1)
p = 8,4: (i,i) (i,i+1) (i,i+2)

p = 9: (i,i+1)
p = 10: (i,i+1) (i,i+2)
p = 11: (i,i+1) (i,i+2)
p = 12: (i,i+1) (i,i+2)

Calculation scheme for exact distances



Found: One conformation found by iBP has an $\mathrm{RMSD} < 2$Å w.r.t. the PDB structure.
Mirror found: One conformation found by iBP has an $\mathrm{RMSD} < 2$Å w.r.t. the mirror image of the PDB structure.
Partially found: At least 50% of the structure of one conformation found by iBP has an $\mathrm{RMSD} < 2$Å w.r.t. the PDB structure.
Not found: None of the previous cases.

Detection of PDB structure if five conformations are generated

Restraint types			i,i+1			i,i+1,i+2	i,i+1	i,i+1,i+2		
PDB	Nres	SecStruct	p=5,4	p=6,4	p=7,4	p=8,4	p=9	p=10	p=11	p=12
1CEY	128	abababababab								
2F05	85	aaaa								
2KSL	51	aaaa								
2KXA	24	aa								
2LJ0	65	bbbbb								
2LVR	30	bba								
2LXZ	32	bbb								
2M5X	40	bbab								
2MC6	73	bbba								
2MDI	56	bbb								
2MGV	65	abbbb								
2MH2	64	abaabb								
2MJ6	90	aabbbb								
2MLA	37	babb								
2MNI	92	baabba								
2MP1	77	bbba								
2MW9	33	bbb								
2MXE	47	bbaab								
2N17	56	bbab								
2N2Q	54	babb								
2RUP	58	bbb								
4BYA	75	aaaa								
4OU0	66	abaabb								
4RBX	32	bbb								

Not found Partially found Found Mirror found

Detection of PDB structure if *chirality* is switched off

Restraint types			i,i+1			i,i+1,i+2	i,i+1	i,i+1,i+2		
PDB	Nres	SecStruct	p=5,4	p=6,4	p=7,4	p=8,4	p=9	p=10	p=11	p=12
1CEY	128	abababababab								
2F05	85	aaaa								
2KSL	51	aaaa								
2KXA	24	aa								
2LJ0	65	bbbbb								
2LVR	30	bba								
2LXZ	32	bbb								
2M5X	40	bbab								
2MC6	73	bbba								
2MDI	56	bbb								
2MGV	65	abbbb								
2MH2	64	abaabb								
2MJ6	90	aabbbb								
2MLA	37	babb								
2MNI	92	baabba								
2MP1	77	bbba								
2MW9	33	bbb								
2MXE	47	bbaab								
2N17	56	bbab								
2N2Q	54	babb								
2RUP	58	bbb								
4BYA	75	aaaa								
4OU0	66	abaabb								
4RBX	32	bbb								

Not found Partially found Found Mirror found

Detection of PDB structure if five conformations are generated

Restraint types			i,i+1			i,i+1,i+2	i,i+1	i,i+1,i+2		
PDB	Nres	SecStruct	p=5,4	p=6,4	p=7,4	p=8,4	p=9	p=10	p=11	p=12
1CEY	128	abababababab								
2F05	85	aaaa								
2KSL	51	aaaa								
2KXA	24	aa								
2LJ0	65	bbbbb								
2LVR	30	bba								
2LXZ	32	bbb								
2M5X	40	bbab								
2MC6	73	bbba								
2MDI	56	bbb								
2MGV	65	abbbb								
2MH2	64	abaabb								
2MJ6	90	aabbbb								
2MLA	37	babb								
2MNI	92	baabba								
2MP1	77	bbba								
2MW9	33	bbb								
2MXE	47	bbaab								
2N17	56	bbab								
2N2Q	54	babb								
2RUP	58	bbb								
4BYA	75	aaaa								
4OU0	66	abaabb								
4RBX	32	bbb								

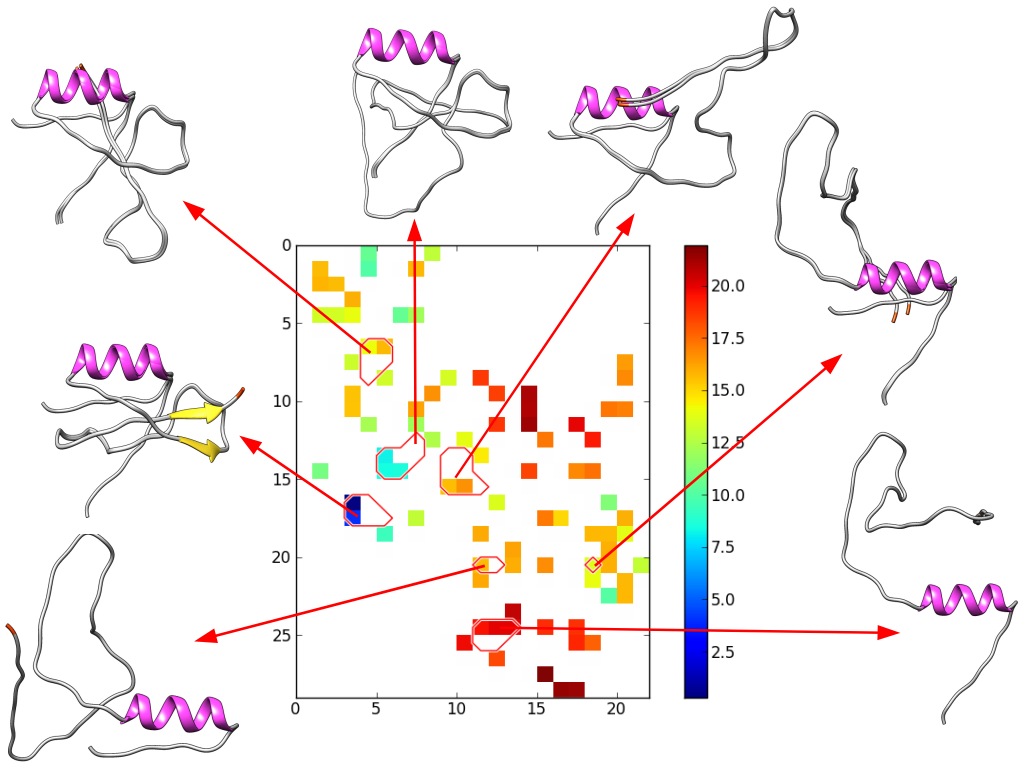
Not found Partially found Found Mirror found

Conformational exploration for short-range restraints

PDB	nbres	SecStruct	rank	RMSD (Å)	Number of saved conformations	Number of parsed leaves	CPU
2F05	85	aaaa	5.37×10^3	0.19	3.75×10^7	5.36×10^8	3d20h
2M5X	40	bbab	1.66×10^3	0.08	4.09×10^3	1.63×10^5	10s
2MGV	65	abbb	5.07×10^2	0.09	5.12×10^2	1.63×10^5	19s
2MXE	47	bbaab	1.92×10^3	0.11	2.04×10^3	3.27×10^3	30s
2N17	56	bbab	2.04×10^5	0.17	10^6	$> 9\times10^6$	6d5h
2N2Q	54	babb	1.77×10^3	0.42	1.79×10^3	2.68×10^8	1d19h
2RUP	58	bbb	22	0.09	64	64	9s
4BYA	75	aaaa	9.99×10^3	0.19	10^4	$> 8\times10^5$	90s
4OU0	66	abaabb	1.10×10^5	0.12	10^6	$> 1\times10^7$	21mn

Pruning distances p=8,4: (i,i) (i,i+1) (i,i+2)

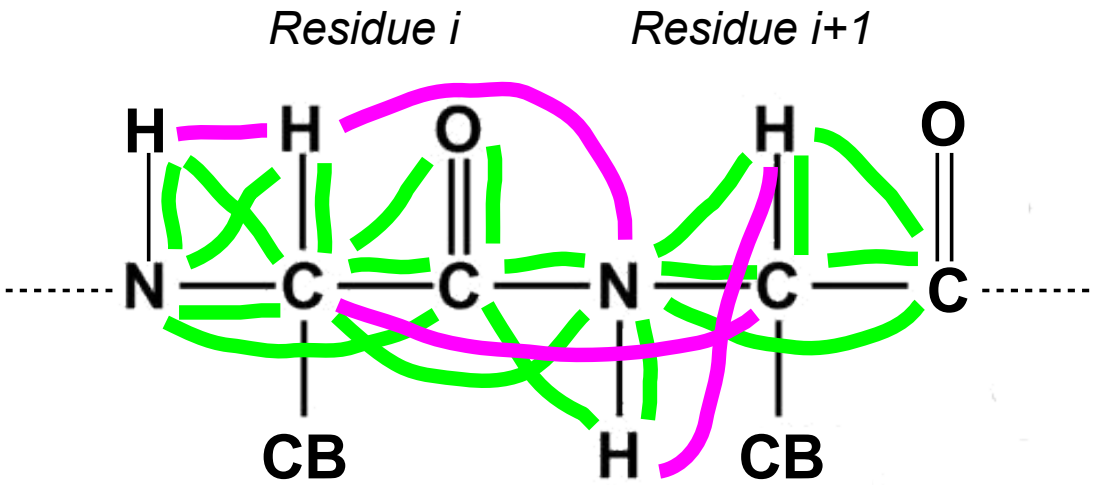
One example of exploration



Bouvier,Desdouits,Ferber,Blondel,Nilges.Bioinformatics 2015

Bouvier,Duclert-Savatier,Desdouits,Meziane-Cherif,Blondel, Courvalin,Nilges,Malliavin.JCIM 2014

Which distance intervals?



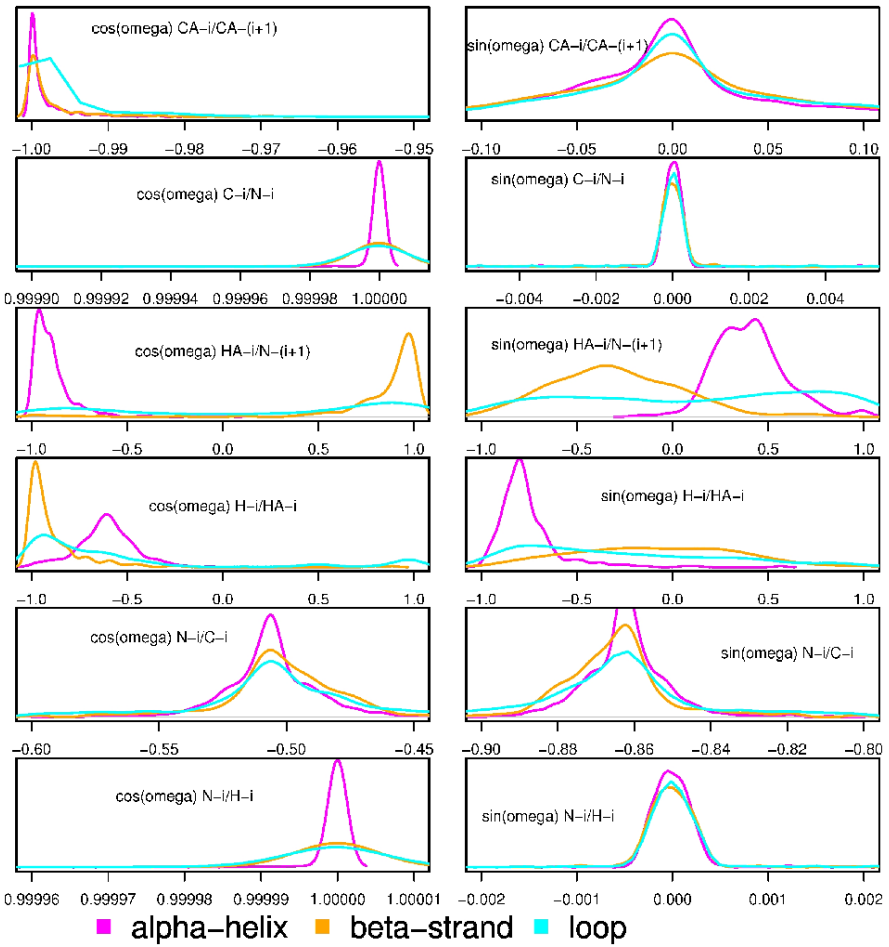
Exact branching distances due to bond and bond angle geometry

Interval branching distances

H-i/HA-i, HA-i/N-(i+1), CA-i/CA-(i+1)

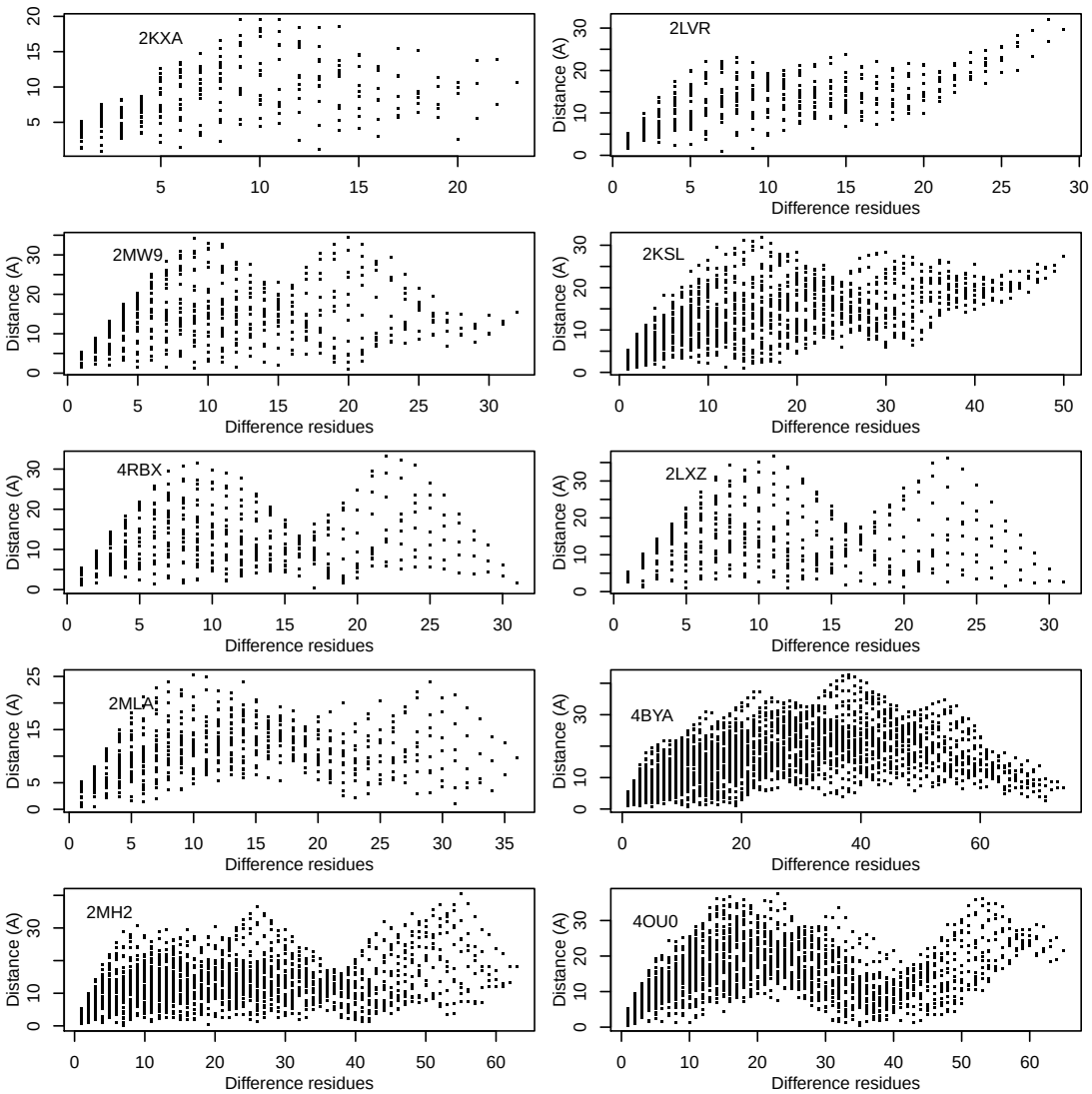
Interval pruning distances between Cα atoms

Strict branching intervals obtained using ω distributions in secondary structure elements

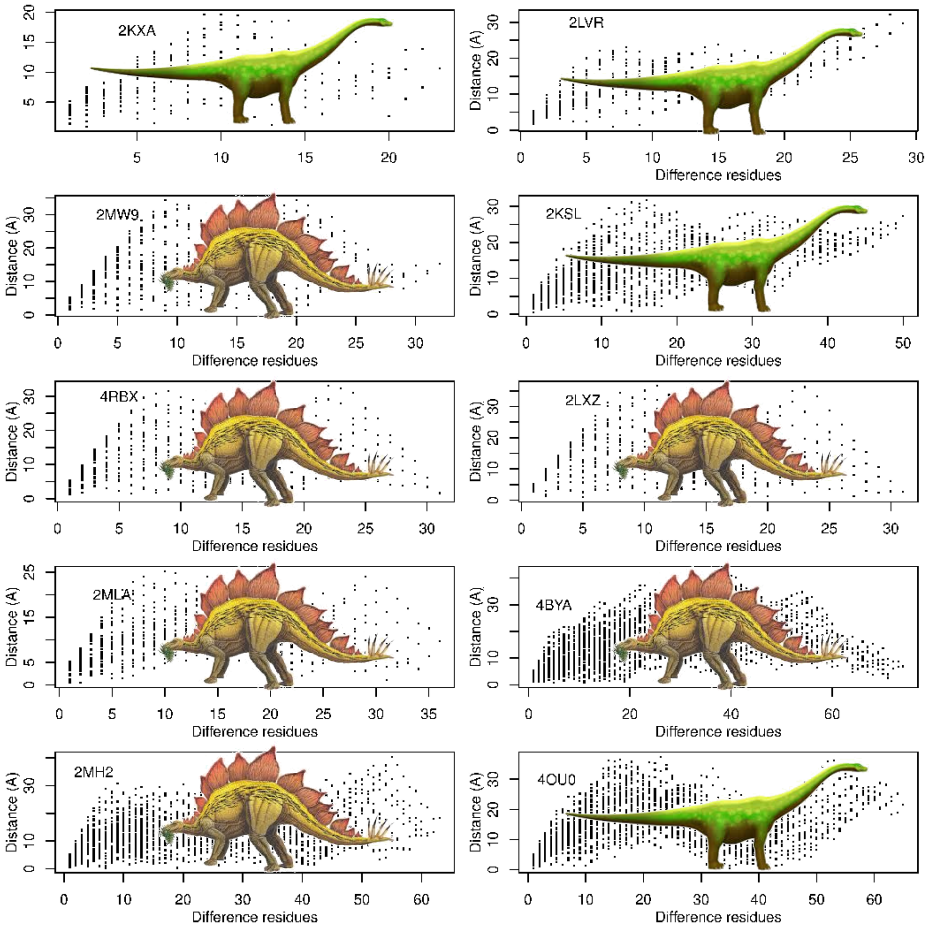


		cos(ω)	sin(ω)
HA-i/N-(i+1)	H	[-1,-0.9]	+
HA-i/N-(i+1)	E	[0.9,1]	+/-
HA-i/N-(i+1)	L	[-1,1]	+/-
H-i/HA-i	H	[-0.75,-0.4]	-
H-i/HA-i	E	[-1,-0.9]	+/-
H-i/HA-i	L	[-1,-0.9]	+/-
CA-i/CA-(i+1)	H	[-1,-0.97]	+
CA-i/CA-(i+1)	E	[-1,-0.97]	+
CA-i/CA-(i+1)	L	[-1,-0.97]	+
N-i/C-i	H	bond	-
N-i/C-i	E	bond	-
N-i/C-i	L	bond	-
N-i/H-i	H	bond	+
N-i/H-i	E	bond	+
N-i/H-i	L	bond	+

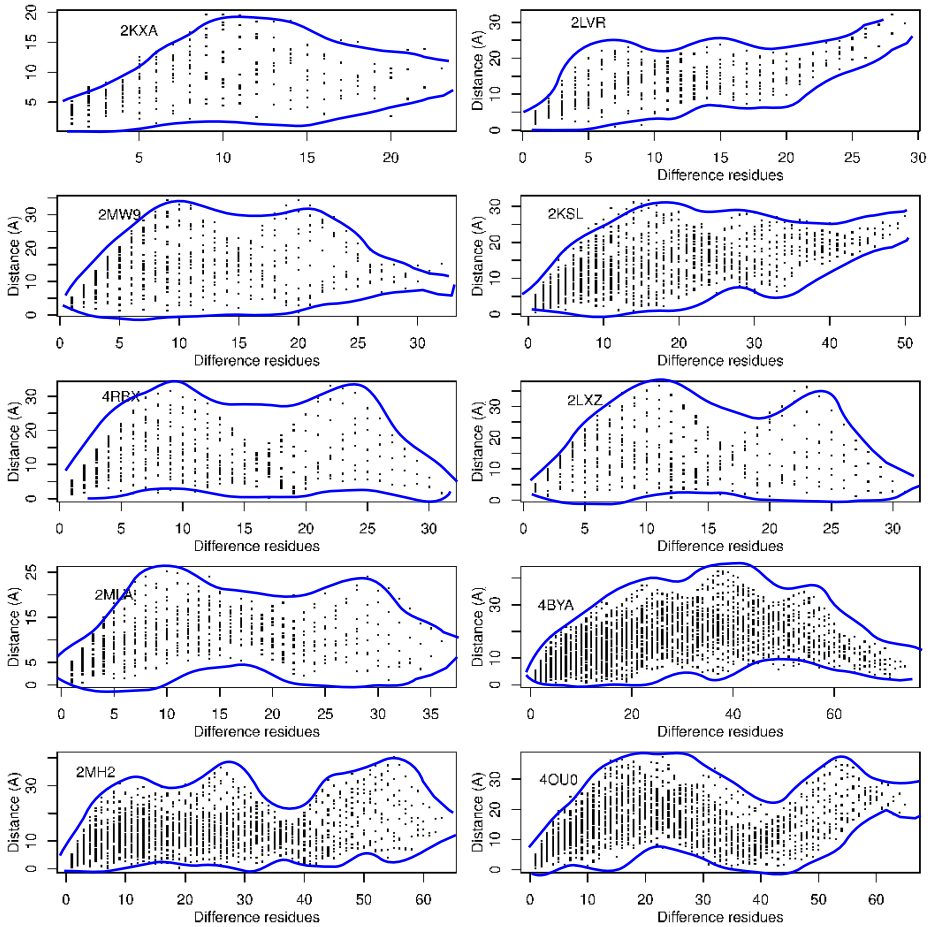
Distribution of distances between atoms Cα



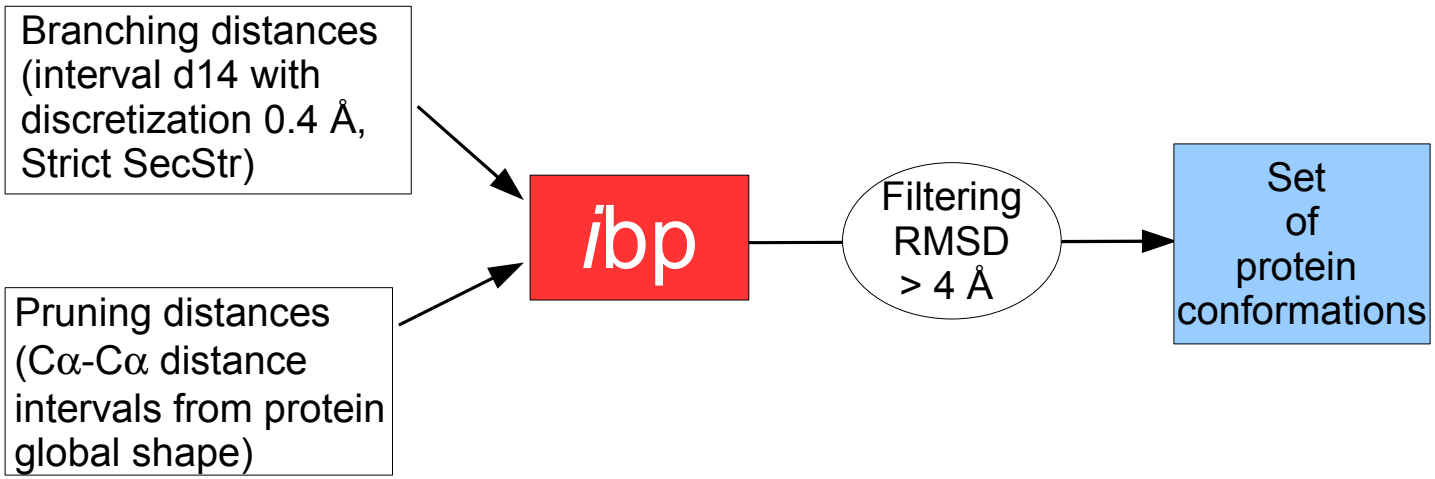
Distribution of distances between atoms Cα



Distribution of distances between atoms Cα



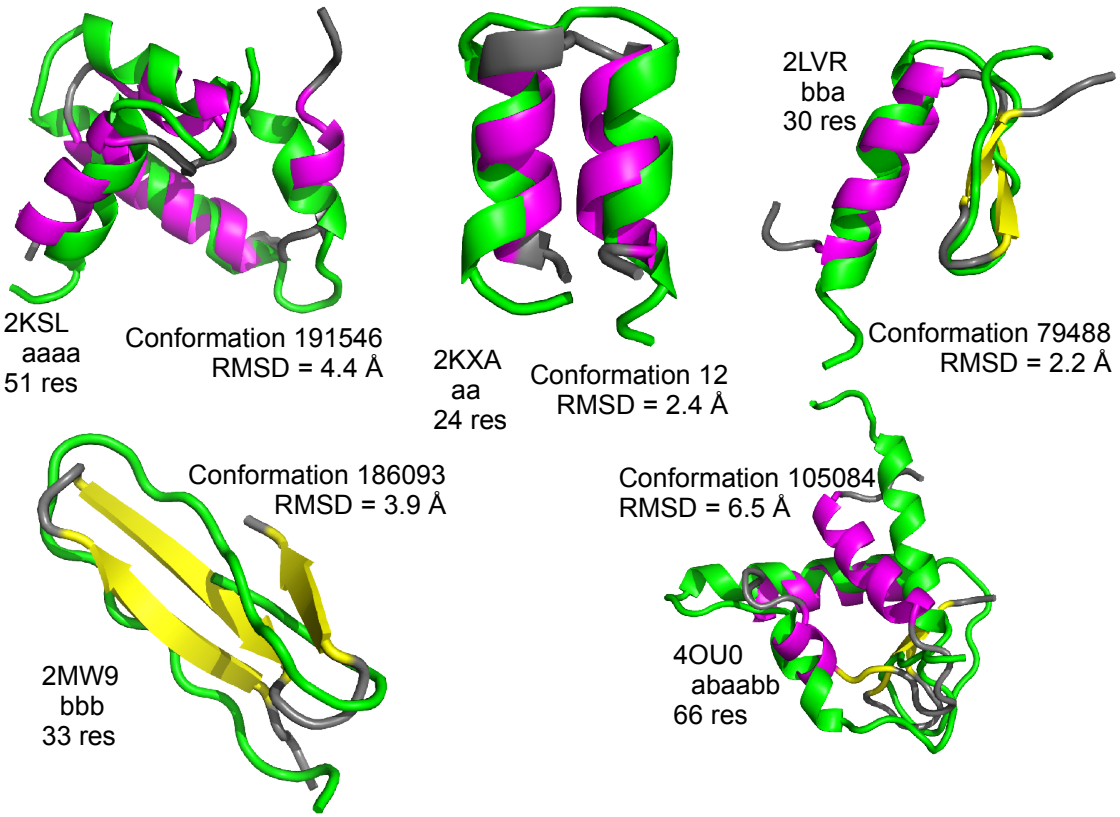
Calculation scheme for intervals



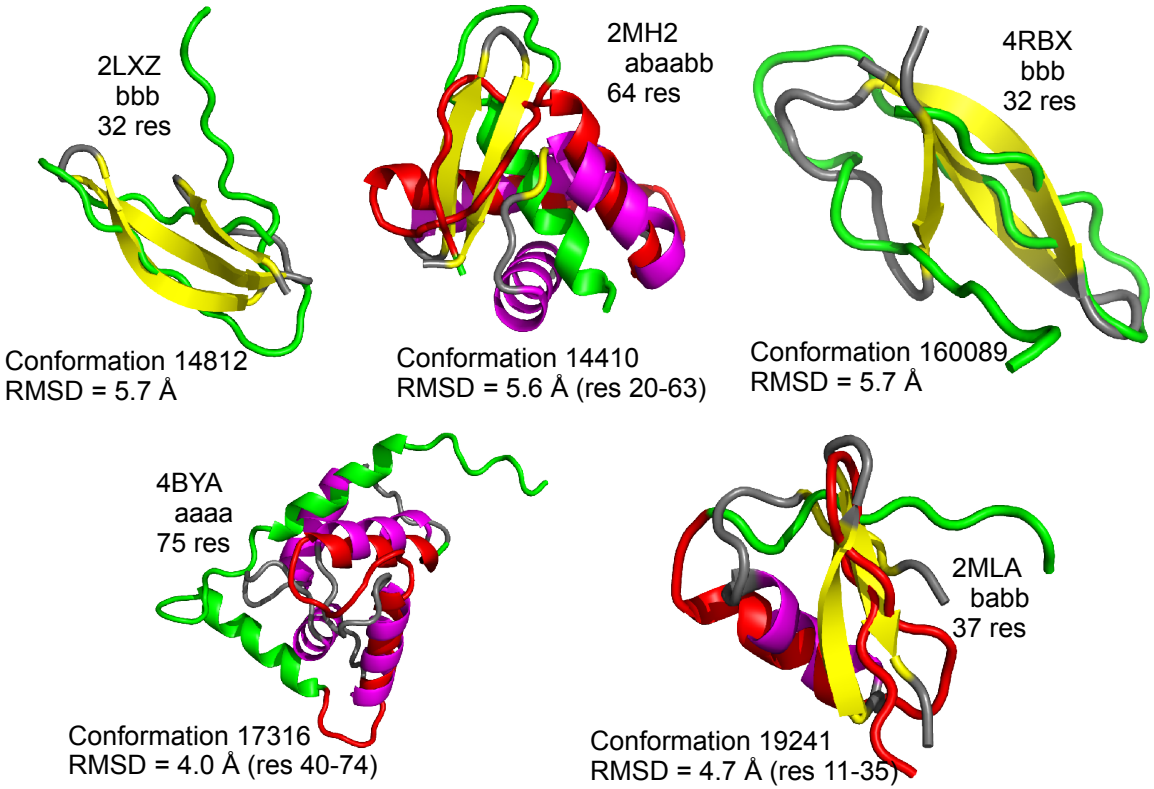
Conformational exploration for short-range restraints

PDB	nbres	SecStruct	rank	RMSD (Å)	Number of saved conformations	Number of parsed leaves	CPU
2KXA	24	aa	12	2.36	**512**	3.27×10^5	1 mn
2LVR	30	bba	7.95×10^4	2.23	200000	2.88×10^8	15h55mn
2MW9	33	bbb	1.86×10^5	3.93	200000	3.14×10^8	11h35mn
2KSL	51	aaaa	1.91×10^5	4.39	200000	8.49×10^7	16h20mn
4RBX	32	bbb	1.60×10^5	4.77	200000	1.84×10^9	4d15h40mn
2LXZ	32	bbb	1.48×10^4	5.69	200000	2.15×10^9	5d8hrs
2MLA	37	babb	1.92×10^4	5.33	200000	2.68×10^9	6d9h50mn
4OU0	66	abaabb	1.05×10^5	6.55	200000	1.64×10^9	6d14h20mn
2MH2	64	abaabb	1.44×10^4	8.48	200000	1.34×10^9	6d1h55mn
4BYA	75	aaaa	1.73×10^4	8.68	200000	1.05×10^7	8h18mn

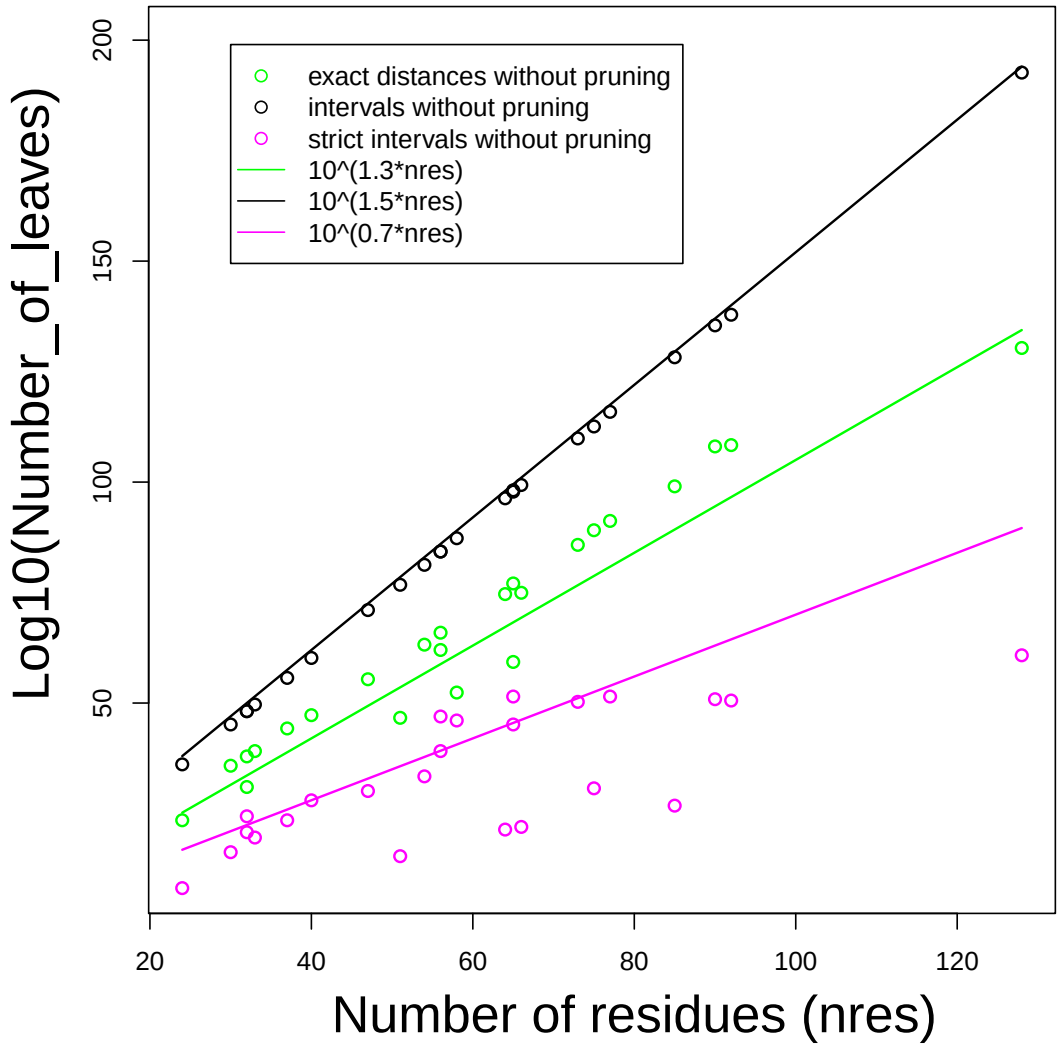
Some examples of calculations with intervals and pruning



Some examples of calculations with intervals and pruning

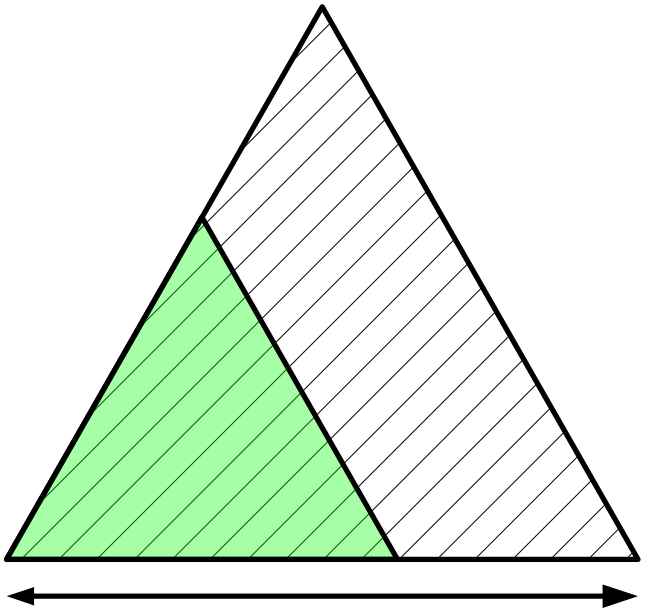


Problem complexity



discretization factor: 0.4 Å

Problem complexity



Number of parsed leaves without pruning = N/NF

Reduction factor due to pruning: F = ns/(N/NF)

Predicted maximum number of leaves with pruning = F. N

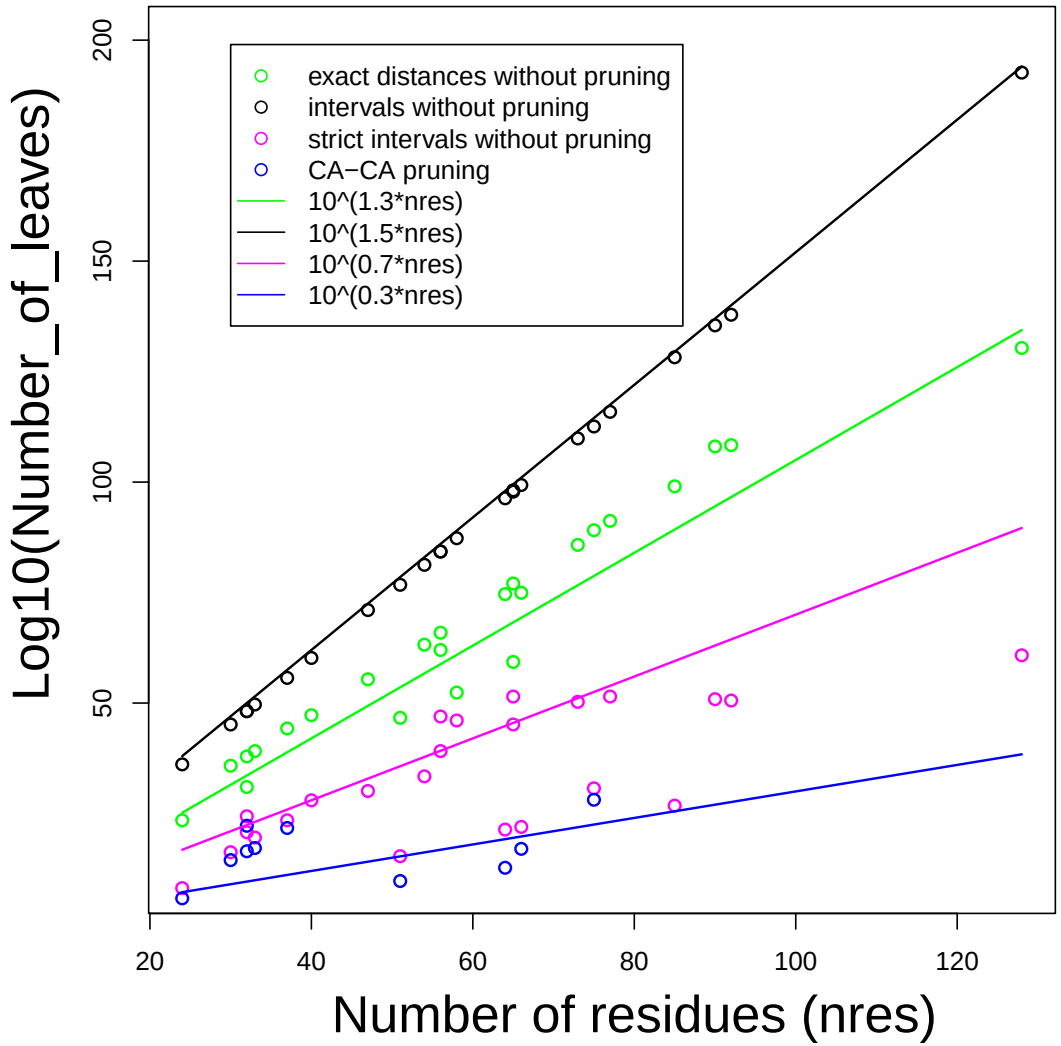
N= maximum number of leaves without pruning

ns= number of parsed leaves using pruning

NF= number of leaves to be parsed without pruning

discretization factor: 0.4 Å

Problem complexity



discretization factor: 0.4 Å

Conclusions and perspectives

- *i*bp provides an algorithm designed for the systematic exploration of the configuration space of protein conformation
- *i*bp is efficient for the exploration of protein conformational space with limited sets of exact short-range inter-atomic distances
- the case of distance intervals could be addressed using stricty definition of secondary structure elements coupled to pruning based on global distance distributions

Acknowledgments

Carlile Lavor UNICAMP

Leo Liberti E. Polytechnique

Benjamin Bardiaux I. Pasteur

Bradley Worley I. Pasteur

Michael Nilges I. Pasteur

Andrea Cassioli E. Polytechnique

Antonio Mucherino U. Rennes

Guillaume Bouvier I. Pasteur

Mohamed Machat I. Pasteur

Funding : ANR bip:bip, ERC BayCells; Pasteur Fondation, CNRS, I. Pasteur,