Announcing the 2015
Keystone Symposia meeting on:

Hybrid Methods in Structural Biology

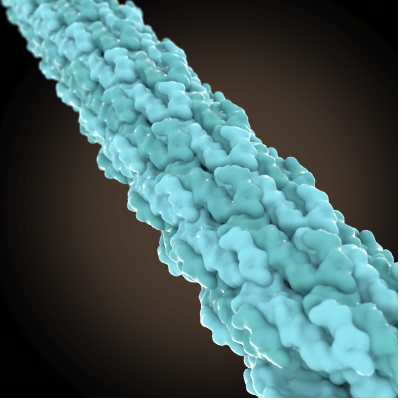
March 4–8, 2015
Granlibakken Resort | Tahoe City, California | USA

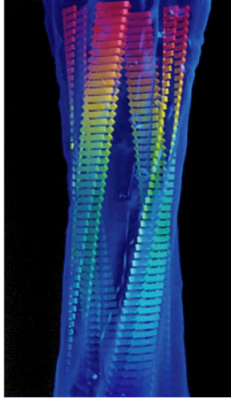
Scientific Organizers:
Jens Meiler, Patrick Cramer** and **Ron A. Milligan

Discounted Abstract/Scholarship Deadline: November 6, 2014
Abstract Deadline: December 9, 2014
Discounted Registration Deadline: January 8, 2015

KEYSTONE SYMPOSIA™
on Molecular and Cellular Biology
Accelerating Life Science Discovery

www.keystonesymposia.org/15C2





CS-Rosetta: A platform for structure determination from sparse NMR data

Workshop

Nik Sgourakis

National Institute for Diabetes and Digestive and Kidney Diseases

National Institutes of Health, USA

KEYSTONE SYMPOSIA™
on Molecular and Cellular Biology
Accelerating Life Science Discovery

Overview

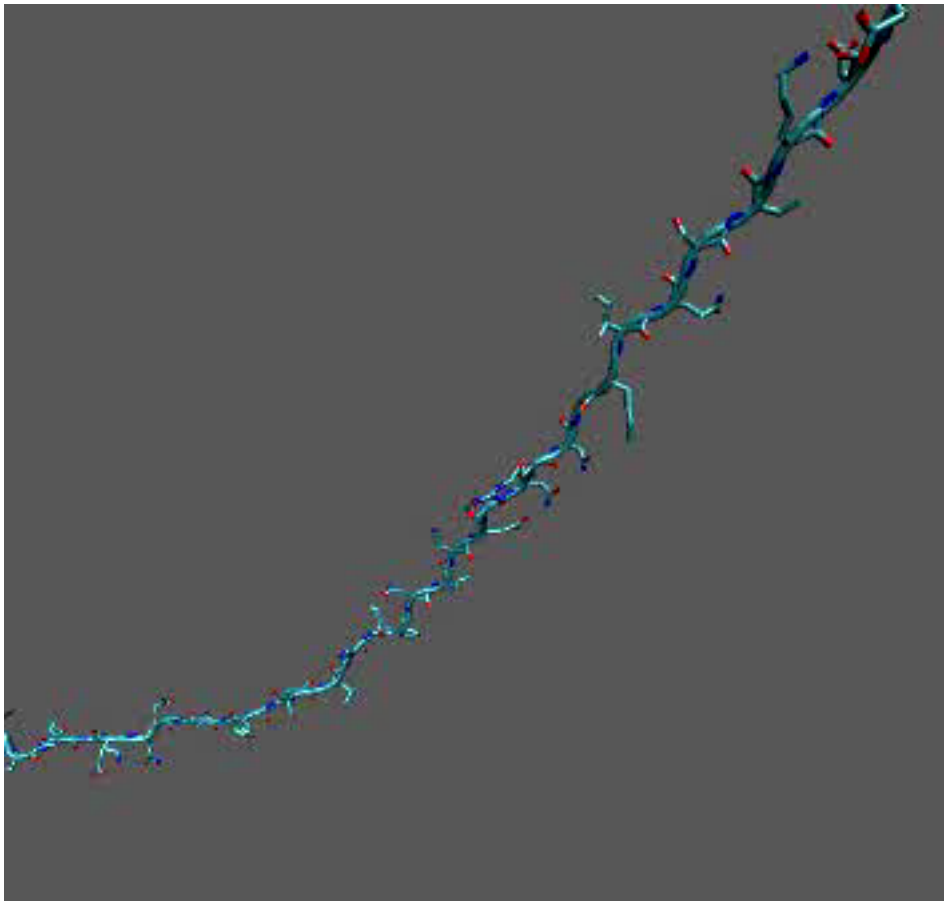
- Basic concepts
- Chemical Shift Rosetta: fragment picking
- CS-Rosetta + Remote homology information
- NOE Restraints
- RDC Restraints
- Symmetric systems
- RosettaOligomers / SAXS data

Conventional NMR structure determination relies on an extensive network of experimental restraints

- 2D, 3D, 4D NMR spectra
- N-H and C-H NOEs
- Chemical shifts
- J-couplings
- Residual Dipolar Couplings

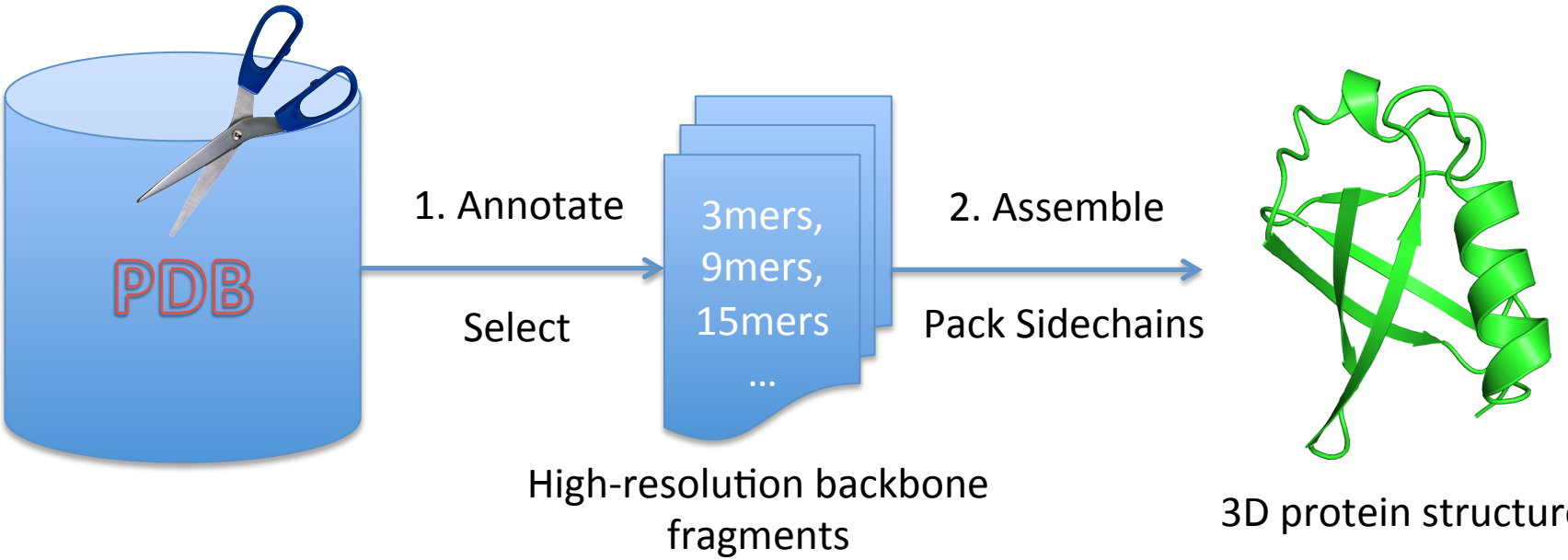
- PREs, SAXS, Relaxation rates..

- Exhaustive Simulated Annealing (XPLOR, ARIA, GROMACS, ..)



$$E_{tot} = \underbrace{E_{chem} + E_{NMR}}_{E_{Bond} + E_{Angle} + E_{Dihedral} + \ldots}$$

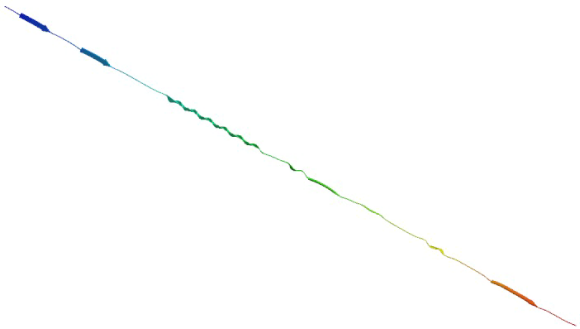
The Molecular Fragment Replacement approach: structure determination using limited NMR data



$$E_{tot} = \cancel{E_{chem}} + E_{NMR}$$

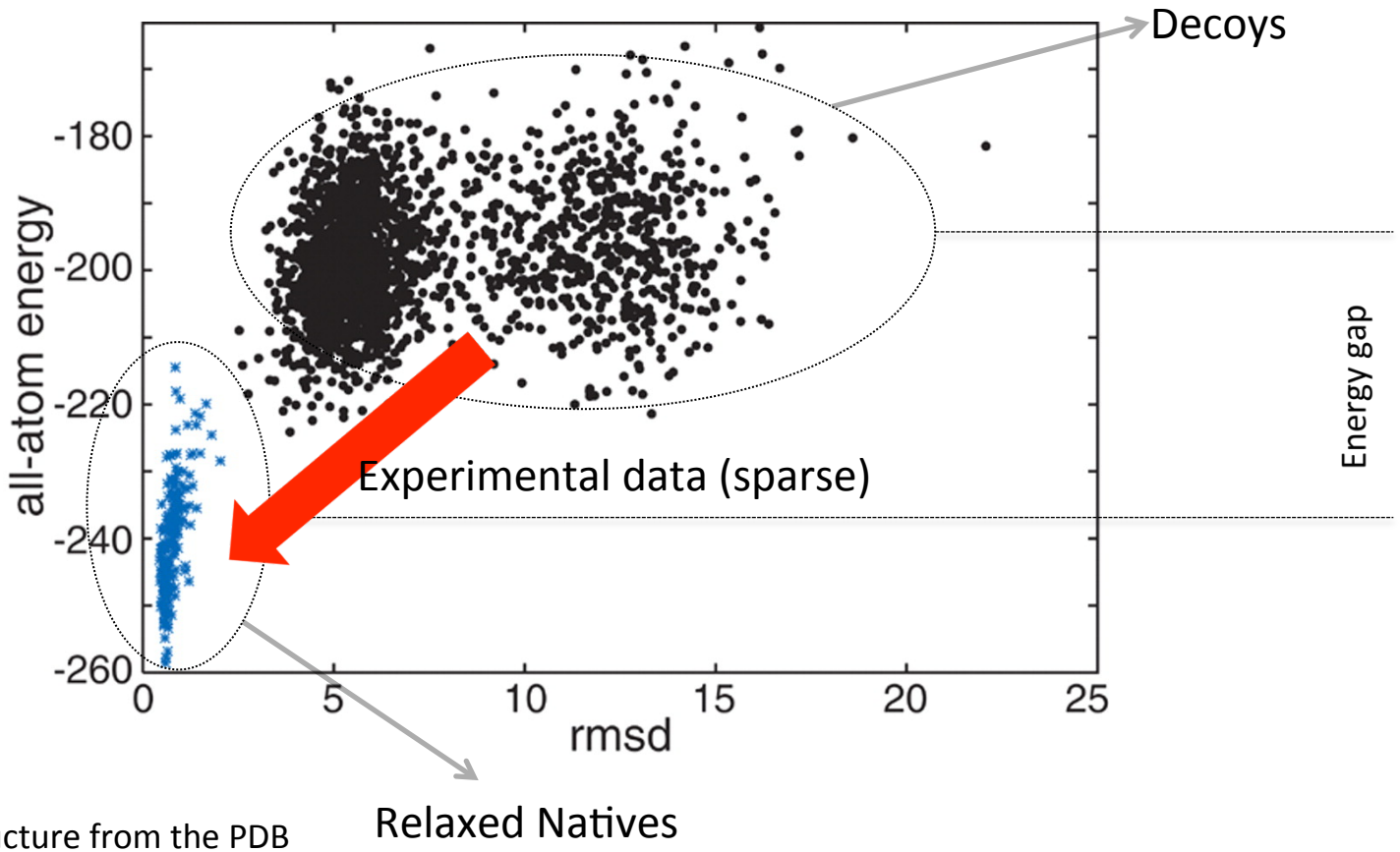
$$E_{tot} = E_{Rosetta} + E_{NMR}$$

$$\underbrace{fa\_rep + rama + hbond + fa\_sol + fa\_dun + fa\_elec + \ldots}_{E_{Rosetta}}$$



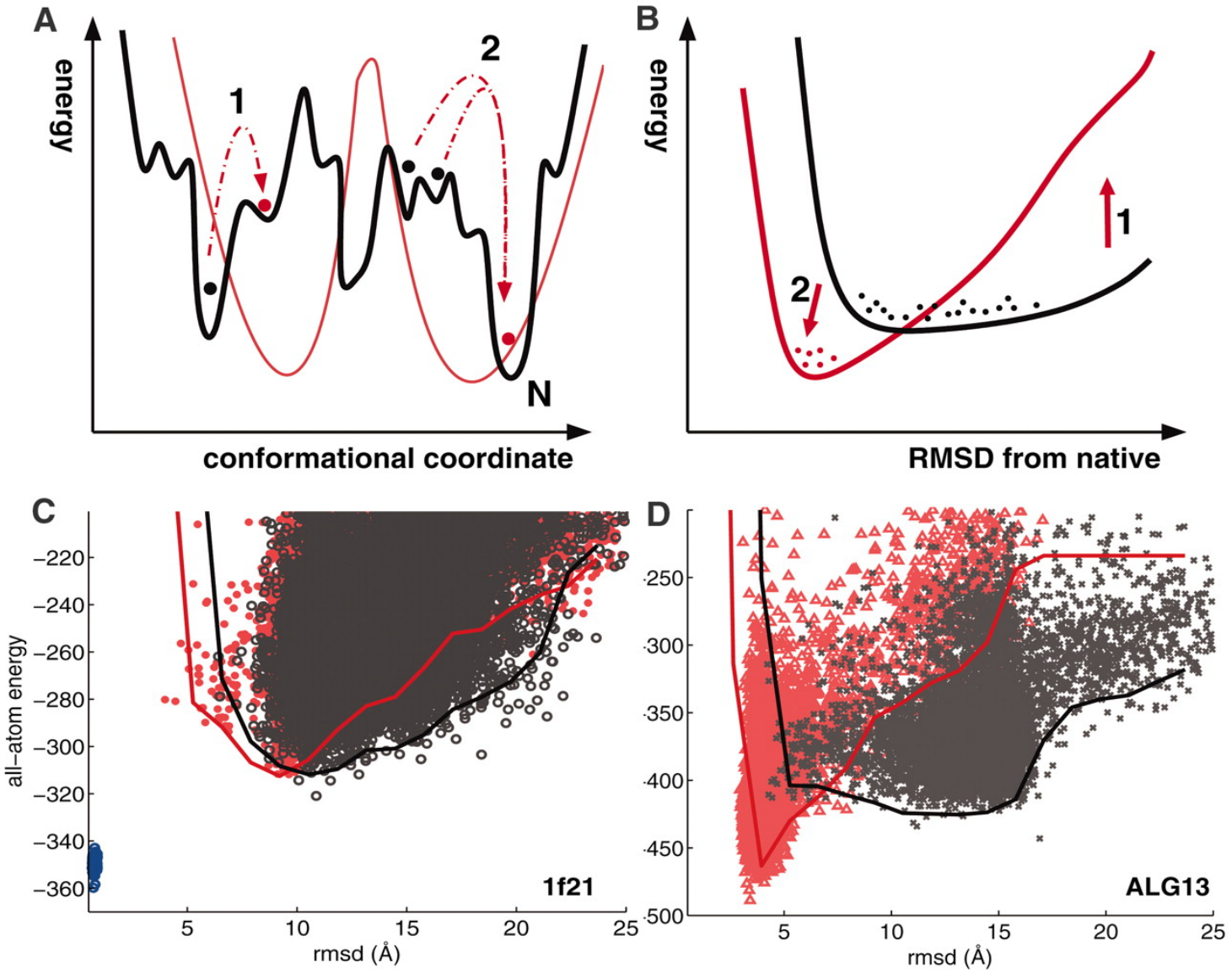
MFR: Frank Delaglio, Ad Bax, JACS, 2000
CHESIRE: Cavalli, Vendruscolo et al., PNAS, 2007
CS-Rosetta: David Baker, Ad Bax, PNAS, 2008

A physically realistic energy function can adequately discriminate the native structure from *de novo* non-native decoys



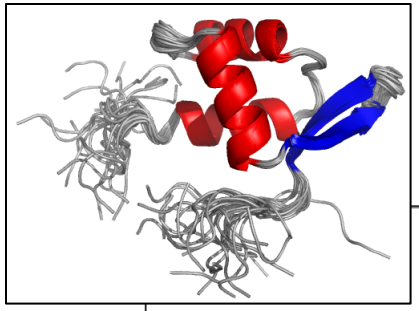
Bradley et al., *Science, 2005*

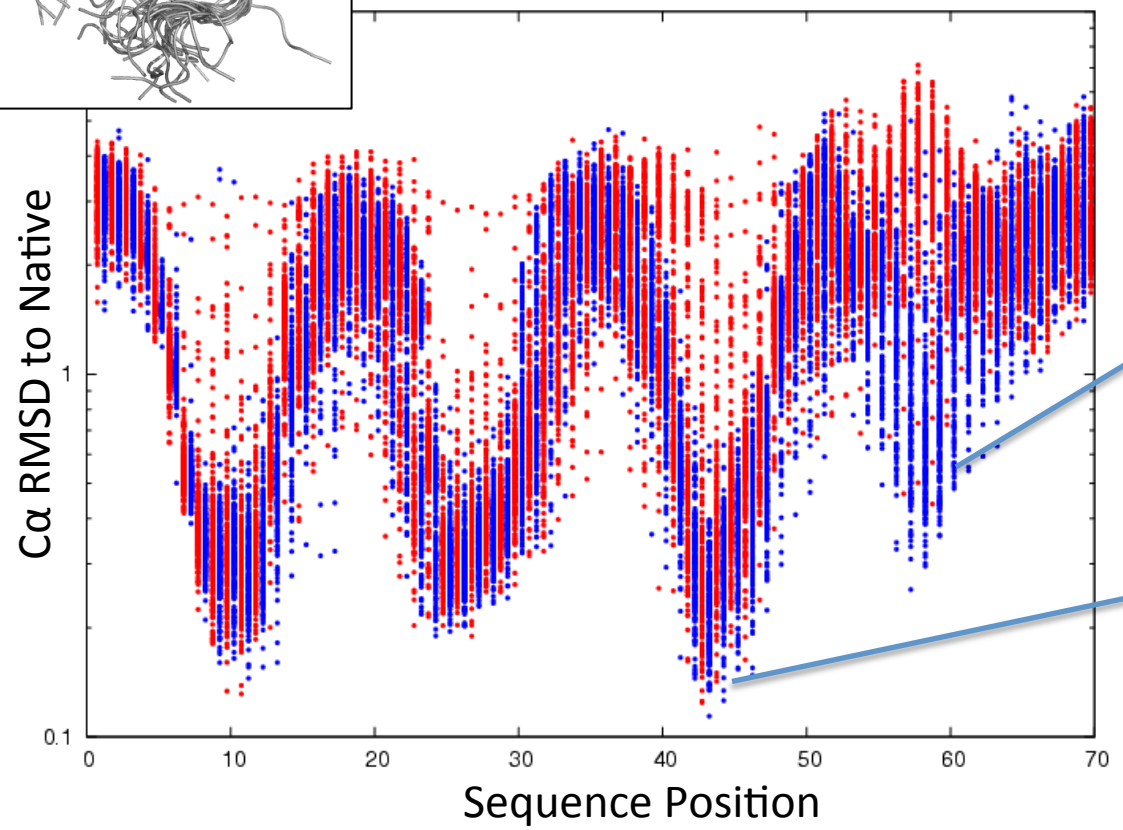
Energy-based methods drastically limit the extent of NMR data required to converge on the native structure

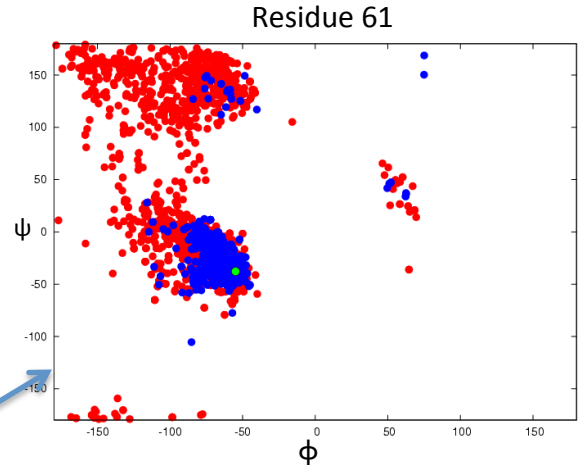


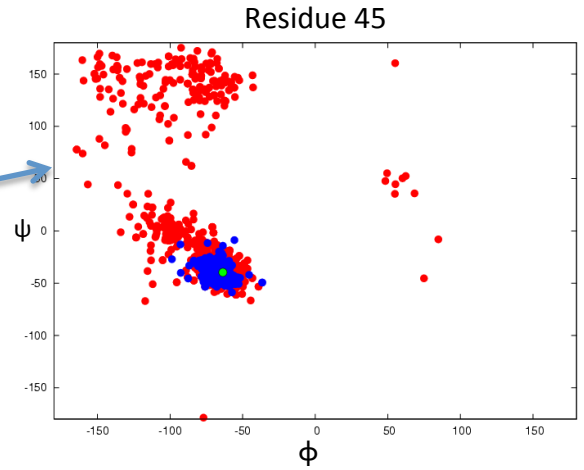
Raman, Lange et al., *Science, 2010*

Chemical shift data reporting on backbone conformation enhance the selection of native fragments









Red: Rosetta Fragments, picked by Sequence profile & Secondary Structure Prediction

Blue: CSRosetta Fragments, picked by CS Comparison & Sequence Matching

Robert Vernon

Chemical shift fragment picking

INPUTS:

TALOS-N secondary structure prediction files (!!! Remove flexible termini and referencing errors!!):

Pred.ss.tab

Pred.tab

BLAST checkpoint file:

blast.checkpoint

Score definition file:

scores.score.cfg

```
# score name        priority  wght   min_allowed  extras
CSScore             400     1      -
ProfileScoreL1       300    1.0     -
TalosSSSimilarity     200    0.25    -     talos
RamaScore           100    1       -     talos
PhiPsiSquareWell        50     0.15    -
```

New fragment picker application:

fragment_picker.default.linuxgccrelease

```
-database      [Rosetta database]
-in::file::vall        [vall databased of PDB fragments]
-frags::n_frags          200
-frags::frag_sizes        3 9
-frags::describe_fragments frags.fsc.score
-out::file::frag_prefix   frags.score
-frags::scoring::config     scores.score.cfg
-in::file::checkpoint     blast.checkpoint
-in:file:talos_cs        cs.tab
-frags::ss_pred          pred.ss.tab talos
-in::file::talos_phi_psi  pred.tab
-frags::sigmoid_cs_A 2
```

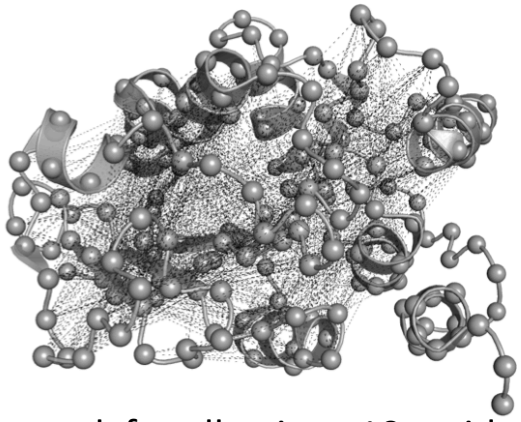
<150 aa targets, simple topologies

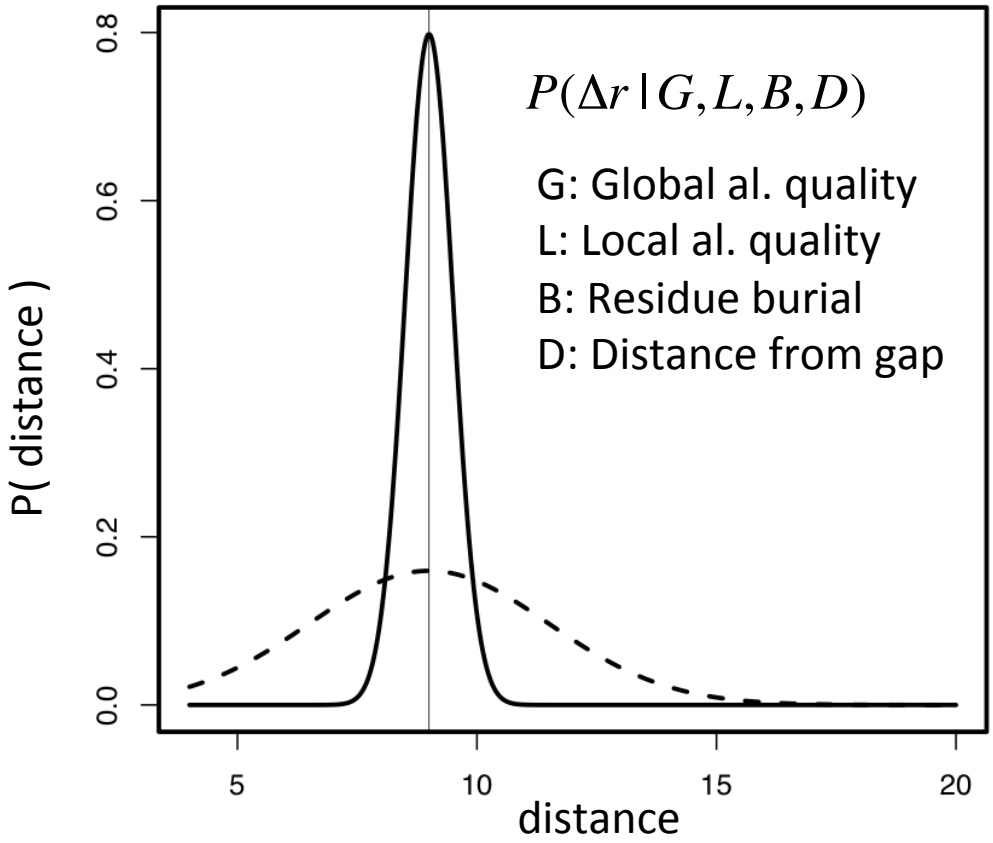
Vernon et al., *JBNMR*, 2013

Comparative modeling basics

```
Query (unknown structure)
LISFATALLKGEDITCFALD

LITFVTAELKGGDITCFGLD
Template Structure (PDB)
```



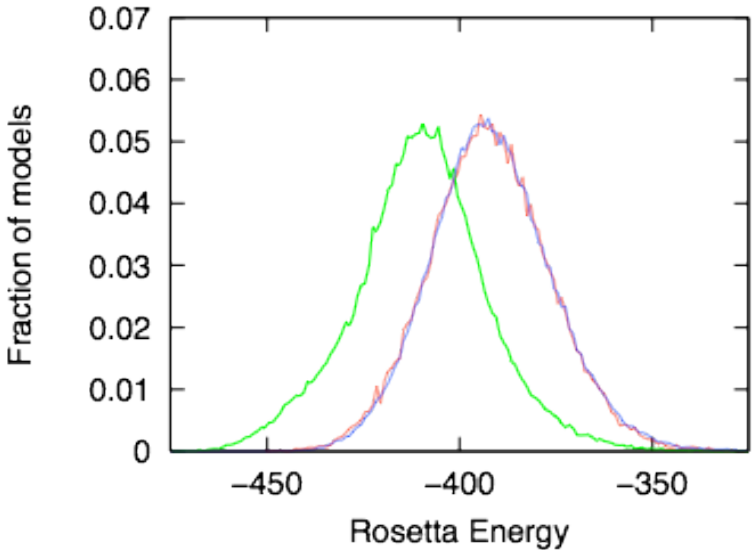


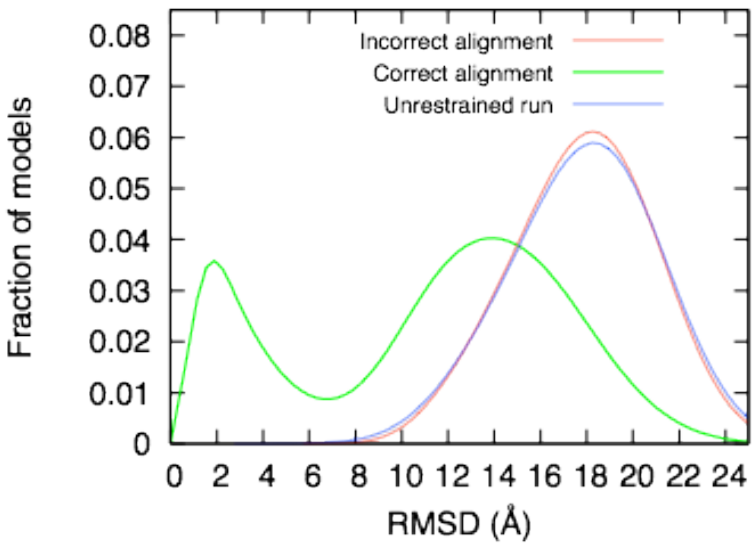
- Solid Gaussian = confident constraint (low entropy)
- Dashed Gaussian = uncertain constraint (high entropy)
- S.D. fitted from 7786 NR PDB structures (10,000 bins)
- Weighted mixture of Gaussians (1 per template)

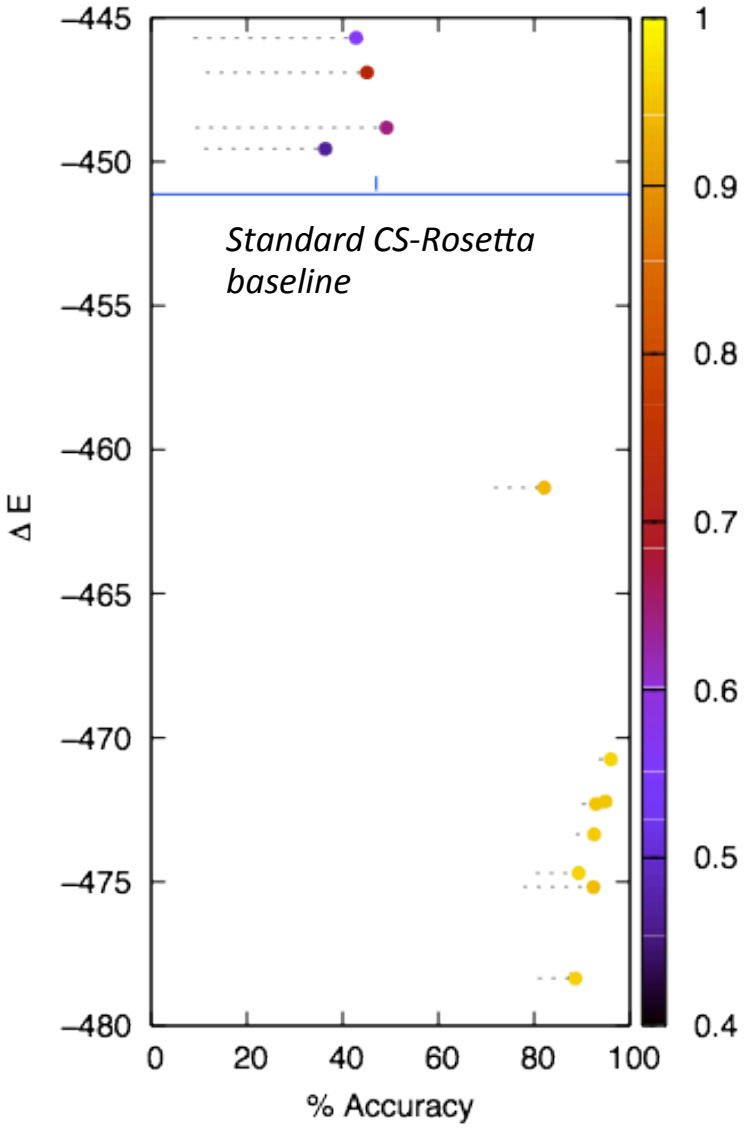
Network for all pairs > 10 residues apart in sequence and < 10 Å in space (otherwise use linear chain model)

Sali and Blundell, *JMB, 1993*
Thompson and Baker, *Proteins, 2011*

CS/HM-Rosetta:
Combining NMR chemical shifts with comparative modeling

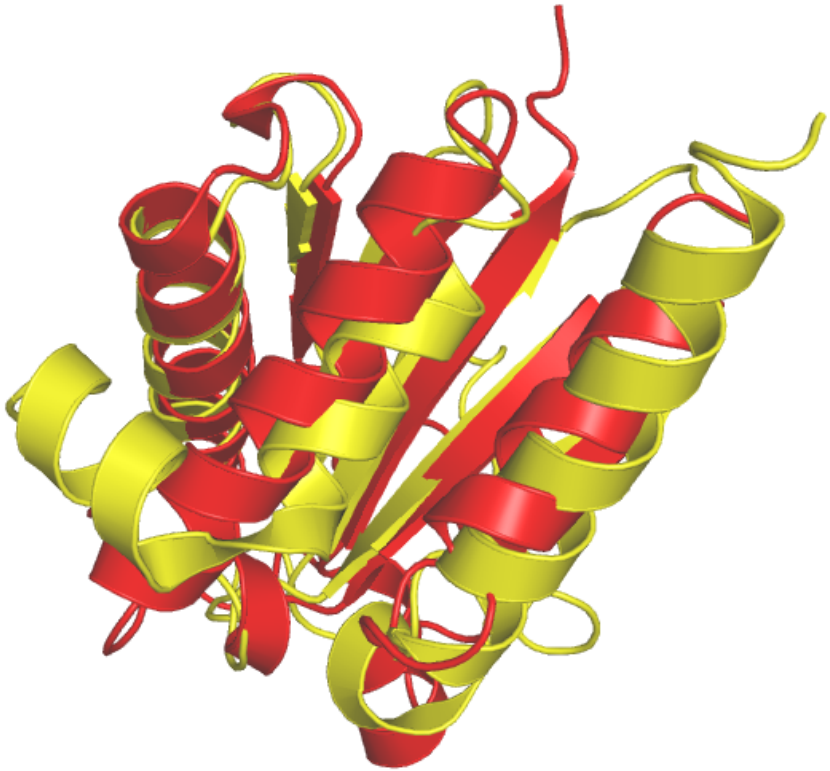




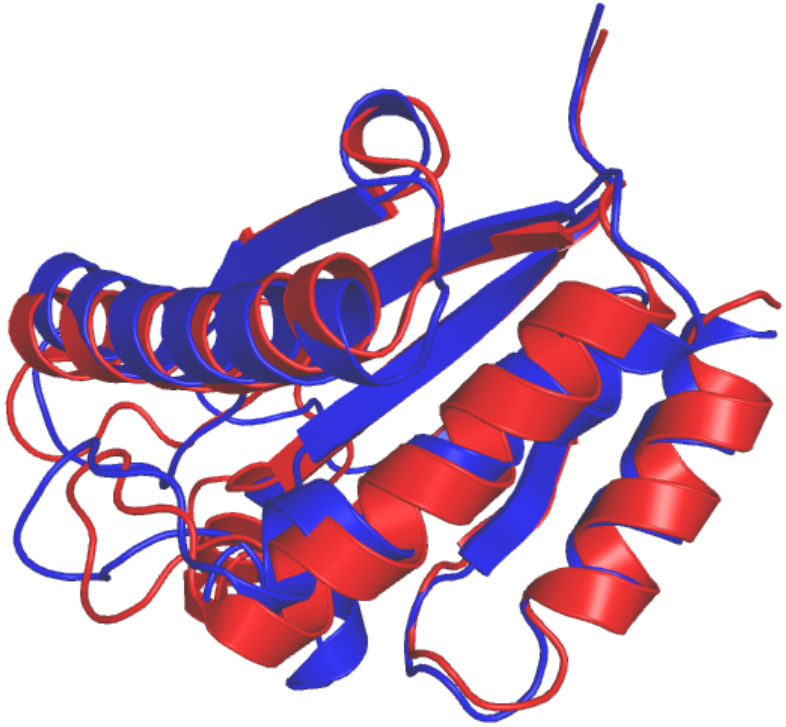


Thompson, Sgourakis et al., *PNAS*, 2012

Adding backbone chemical shift information improves comparative models < 20 % sequence identity

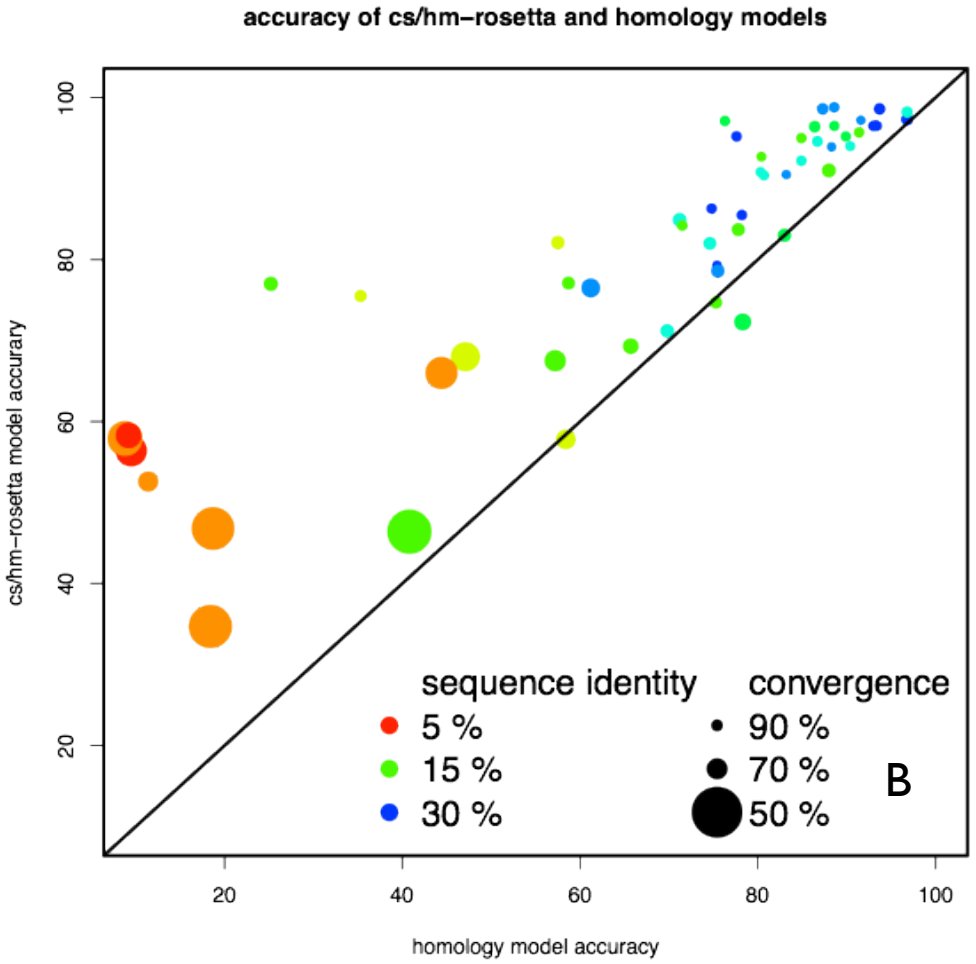


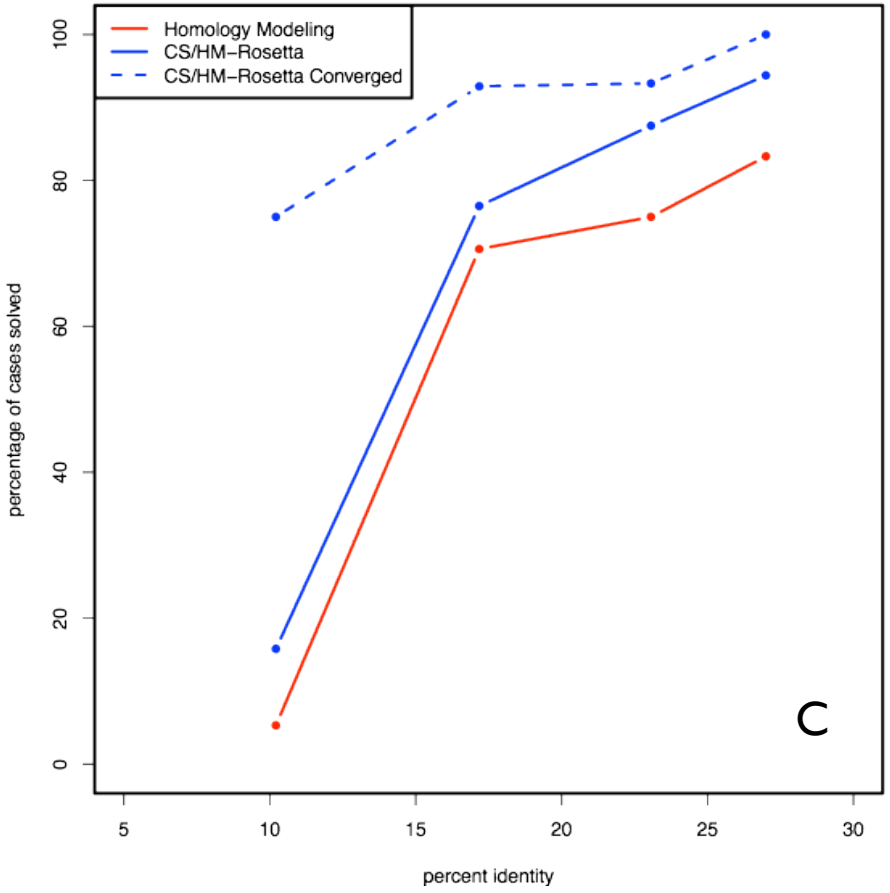
Homology model – 3.5 Å
17% sequence identity



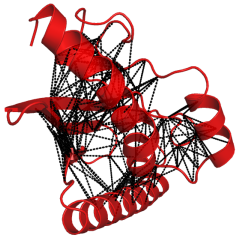
CS/HM-Rosetta – 0.9 Å

CS/HM-Rosetta improves the performance of comparative modeling methods for twilight alignments



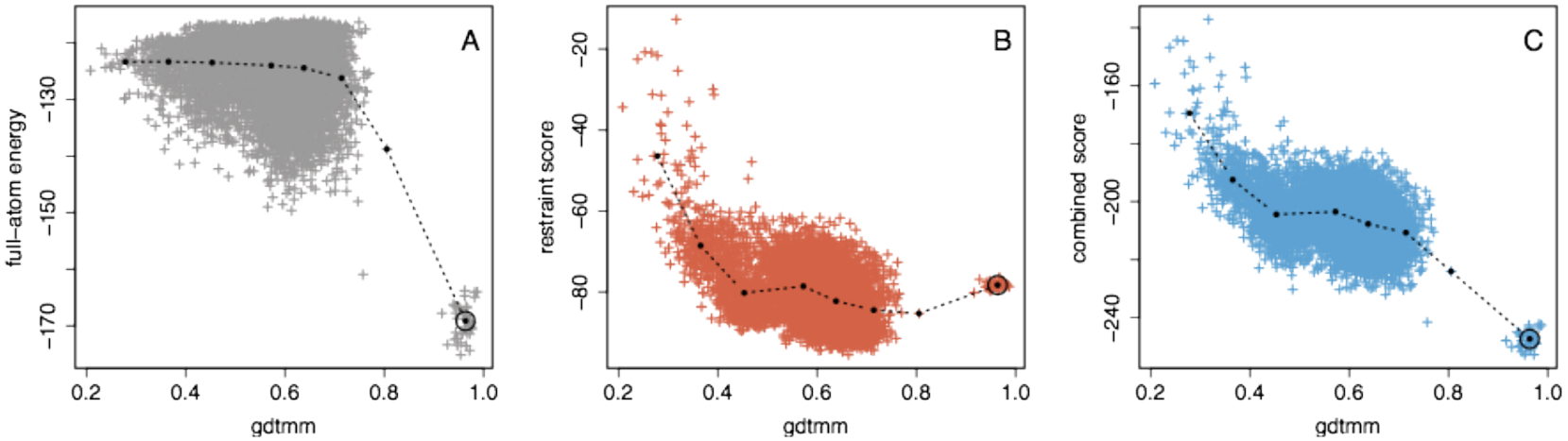


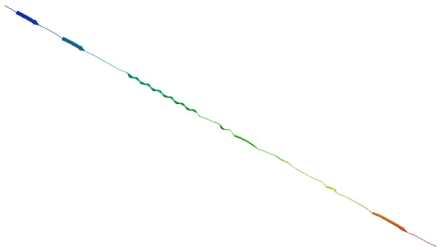
Thompson, Sgourakis et al., *PNAS*, 2012



predict_distances.pl -aln_format grishin [sequence alignments].ali [your sequence].fasta
Output:
 # used 1 alignments from 4f8t.aln.
 # alignment to 4f8tA ln_e_value = -2.30258509299405)
AtomPair CA 26 CA 47 SOGFUNC 1 8.437 3.616 1.000
AtomPair CA 26 CA 48 SOGFUNC 1 5.835 3.616 1.000

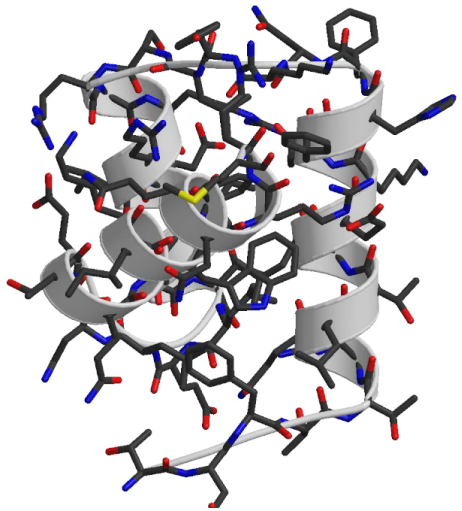
The sparse data eliminate large regions of conformational space during the low resolution sampling stage



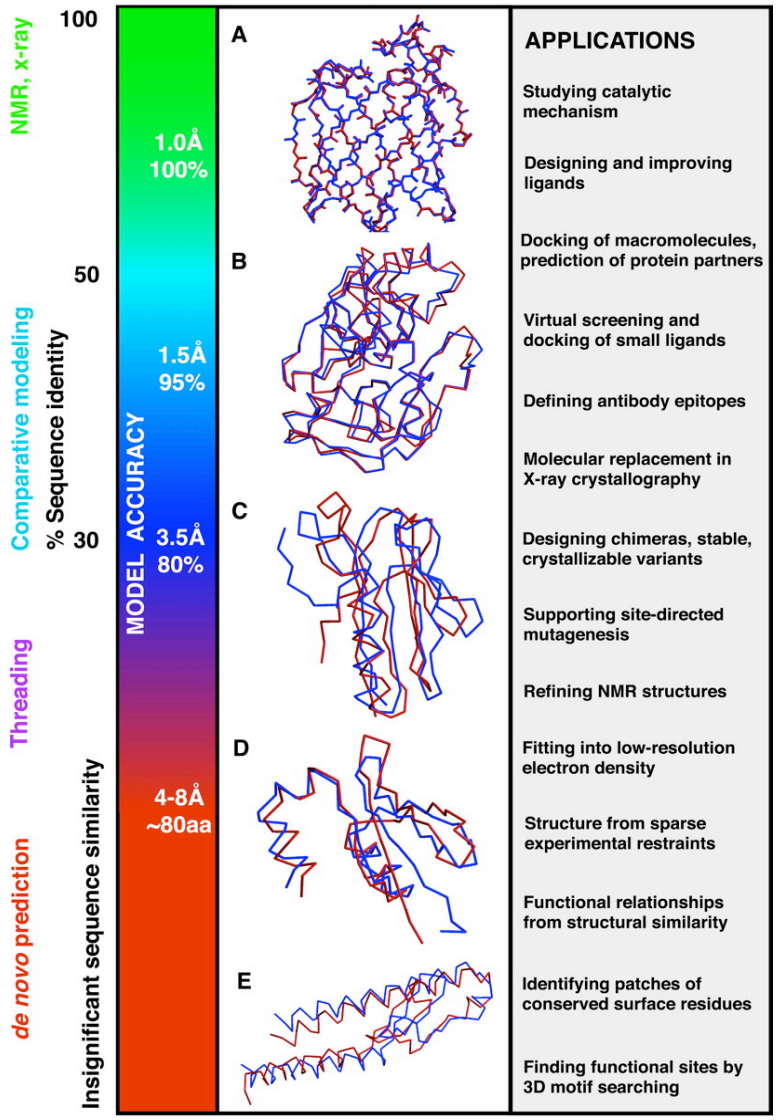


Backbone folding

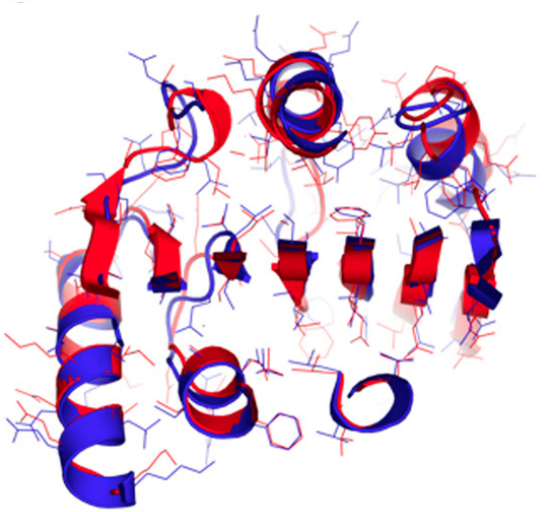
Sidechain packing



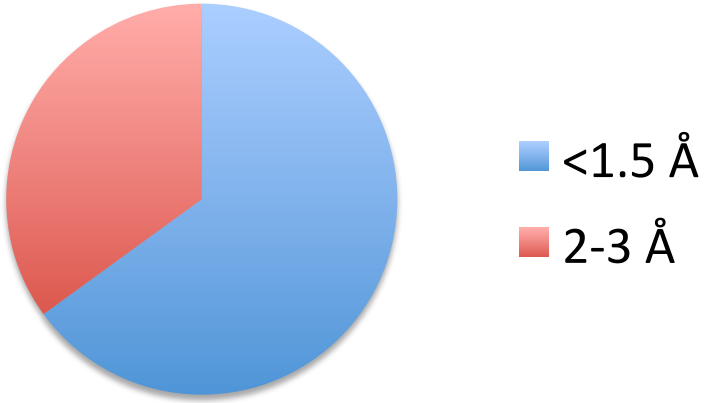
How good are our CS/HM-Rosetta models?



Baker and Sali, *Science*, **2001**



202 residues / 1.3 Å



**Benchmark of 20 pairs of Xray/NMR structures*

NMR structure determination using sparse ILV ^{13}C labeling

Backbone assignment and
^{1}H-methyl distances

↓

NOE distance restraints

↓

CS - Rosetta ← Distance restraints
Chemical shifts
RDCs

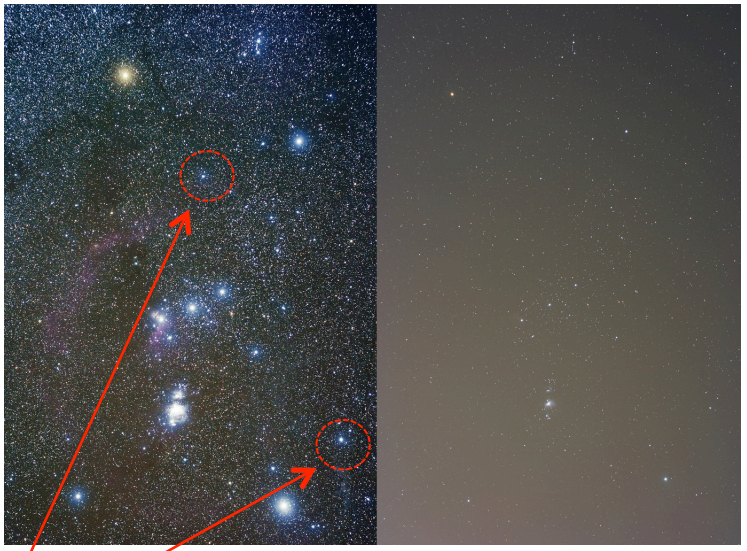
↓

Structure converged?

↓ Yes

Final ensemble

No → Structure converged?



Lange, Rossi, Sgourakis et al., *PNAS*, 2012

Implementation of NOE distance constraints with methyl ambiguities

Simple AtomPair contraints:

AtomPair H 95 H 105 BOUNDED 1.5 3.650 0.3 NOE; amide-amide

Proton Ambiguities:

AmbiguousNMRDistance H 56 QD1 71 BOUNDED 1.5 4.000 0.5 NOE; amide-methyl

Proton/Carbon Ambiguities:

AmbiguousNMRDistance H 55 QQG 94 BOUNDED 1.5 4.000 0.5 NOE; amide-methyl

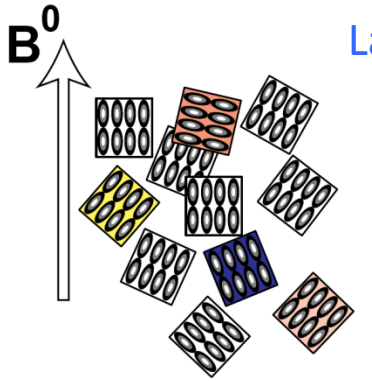
AmbiguousNMRDistance QQD 25 QQG 108 BOUNDED 1.5 4.000 0.5 NOE; methyl-methyl

Mapping into the centroid representation (1.5Å padding per C):

cat [fullatom Rosetta constraint file] | perl map_csts_to_centroid_simple.pl > [centroid Rosetta constraint file]

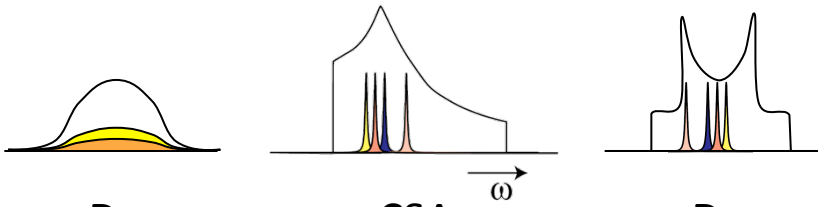
$$E(x) = \begin{cases} 0, & x \geq ll \cap x \leq ul \\ 2^{(x-ul)/0.3}, & x > ul \\ 2^{(ll-x)/0.3}, & x < ll \end{cases}$$

Residual Dipolar Coulings: sensitive probes of protein structure and dynamics



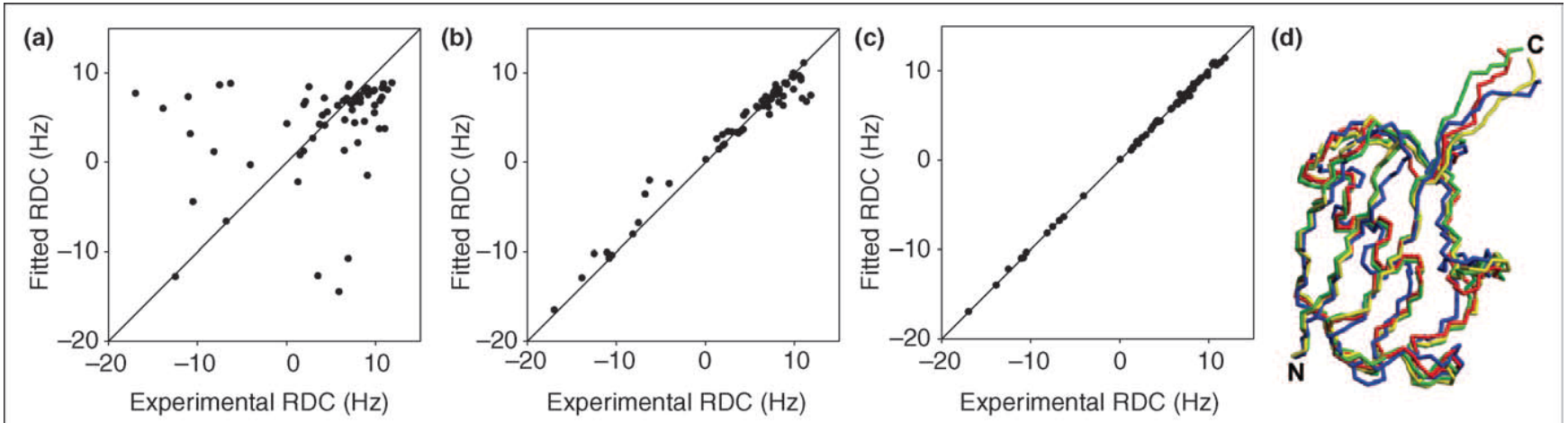
Lack of molecular tumbling

Distribution of crystallite orientations



D_{II} CSA D_{IS}

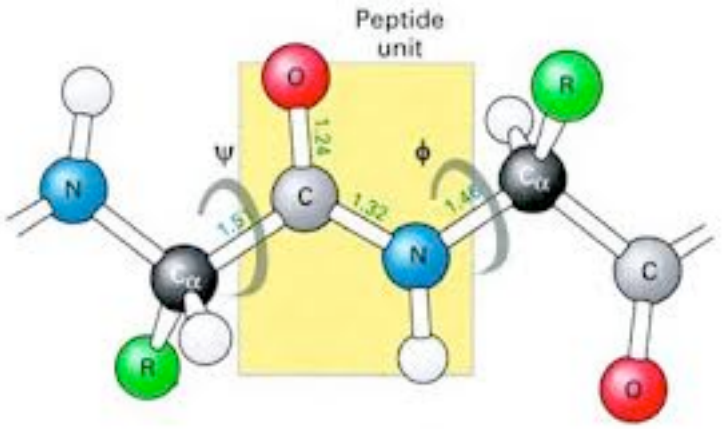
20kHz ^{1}H Dipolar Interaction:
Isotropic solution-> 0Hz splitting
0.1% alignment ->20Hz splitting

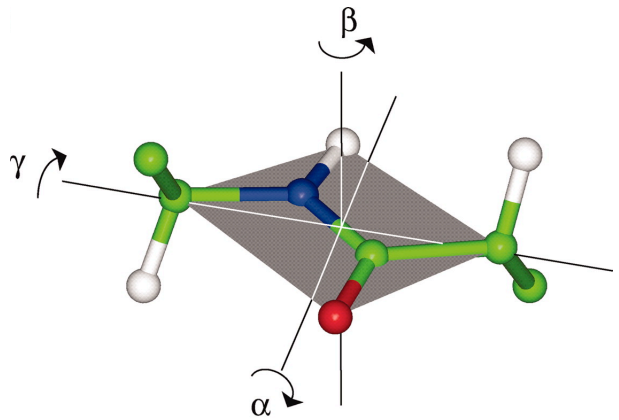


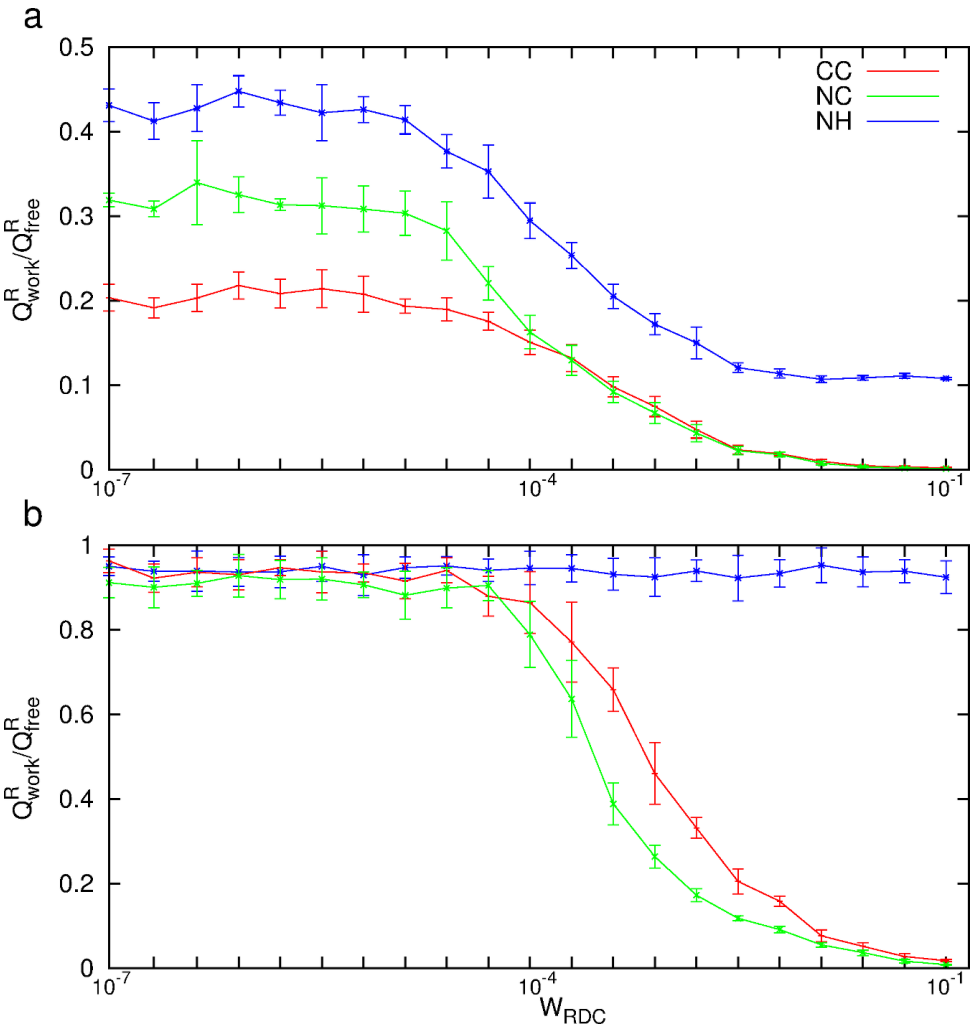
"scrambled" RDCs can be refined to near experimental noise levels!

Current Opinion in Structural Biology
Grishaev and Bax, 2005

Rosetta refinement using multiple RDC datasets converges on the global energy minimum while allowing for cross-validation







Sgourakis et al., *In Preparation*

Implementation of RDC restraints in Rosetta

Multiple bond vector support (data are automatically scaled relative to N-H)

3 N 3 H 6.64

3 C 4 N 3.34

3 C 3 CA 3.4

Support of multiple alignment datasets, scaled by 1/Da

-in:file:rdc *gel.rdc phage.rdc*

-rdc:fix_normAzz W1 W2 (W1, W2 estimated from the input data using the histogram method: histo.py [rdc file] 5)

2 types of scoring modes:

-rdc:fit_method [svd, nls]

-rdc:fixDa [value]

-rdc:fixR [value]

Command-line application for the evaluation of Q-factors, al. tensor parameters calculated RDCs:

r_score_rdc.default.macosgccrelease

-in:file:s [input PDB file]

-in:file:rdc [Rosetta RDC constraint file]

-out:level 999

-rdc:print_rdc_values calc.rdc *|grep 'Da \|| R: |/ Qbax'*

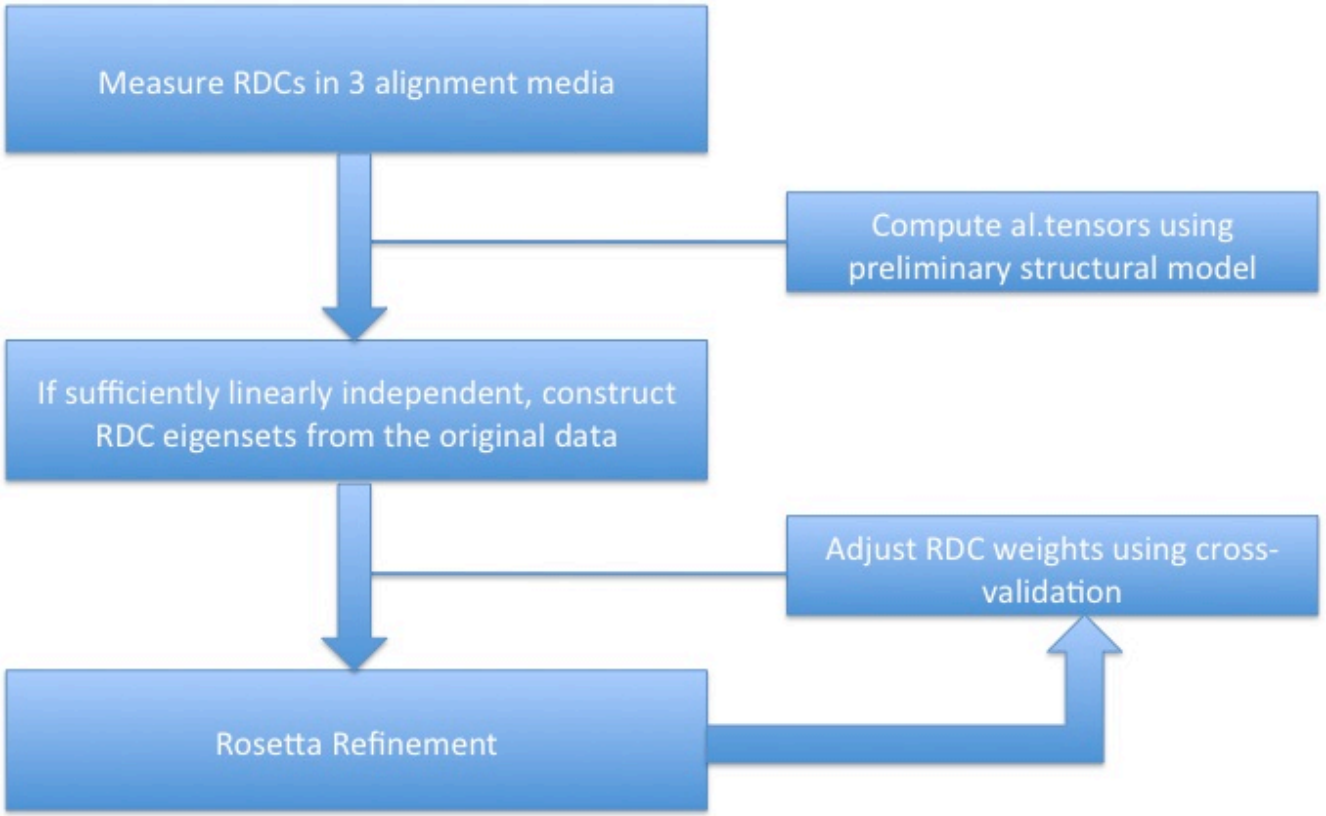
$$Q = \frac{RMS(D_{calc} - D_{obs})}{\sqrt{D_a^2(4+3R^2)/5}}$$

Lei Shi

exactly the same output as PALES

Rossi, Shi, .. , Sgourakis, *Proteins* 2015

Independent cross-validation using multiple alignment data



Generation of Orthogonal alignment datasets

Convert a set of input Saupe parameters into its irreducible representation:

cat [Saupe matrix file] |perl convert_saupe.pl |perl raw2inp.pl

This step generates output files 1.inp, 2.inp, etc. to be used in the next step.

Compute the normalized scalar product between pairs of al. tensors:

cat [Saupe matrix file] | perl dot_tensors.pl

Compute orthogonal alignment tensors using Singular Value Decomposition and extract linear recombination coefficients (Lorieau et al., JBNMR, 2013) :

python rdc_pipe2xplor.py |tail -3 > [linear coefficients file]

Create orthogonally recombined datasets from linear combinations of the original RDC data (implemented for 3 datasets):

perl creat_eigensets.pl [linear coefficients file] [rdc file 1] [rdc file 2] [rdc file 3]

This scripts generates 3 O.R. datasets: eigen1.rdc, eigen2.rdc, eigen3.rdc, in Rosetta format. The input is also in Rosetta format (see below).

The fitted Saupe matrix parameters used in step (5) can be copied directly from the output of the DC program: http://spin.niddk.nih.gov/bax/nmrserver/dc/

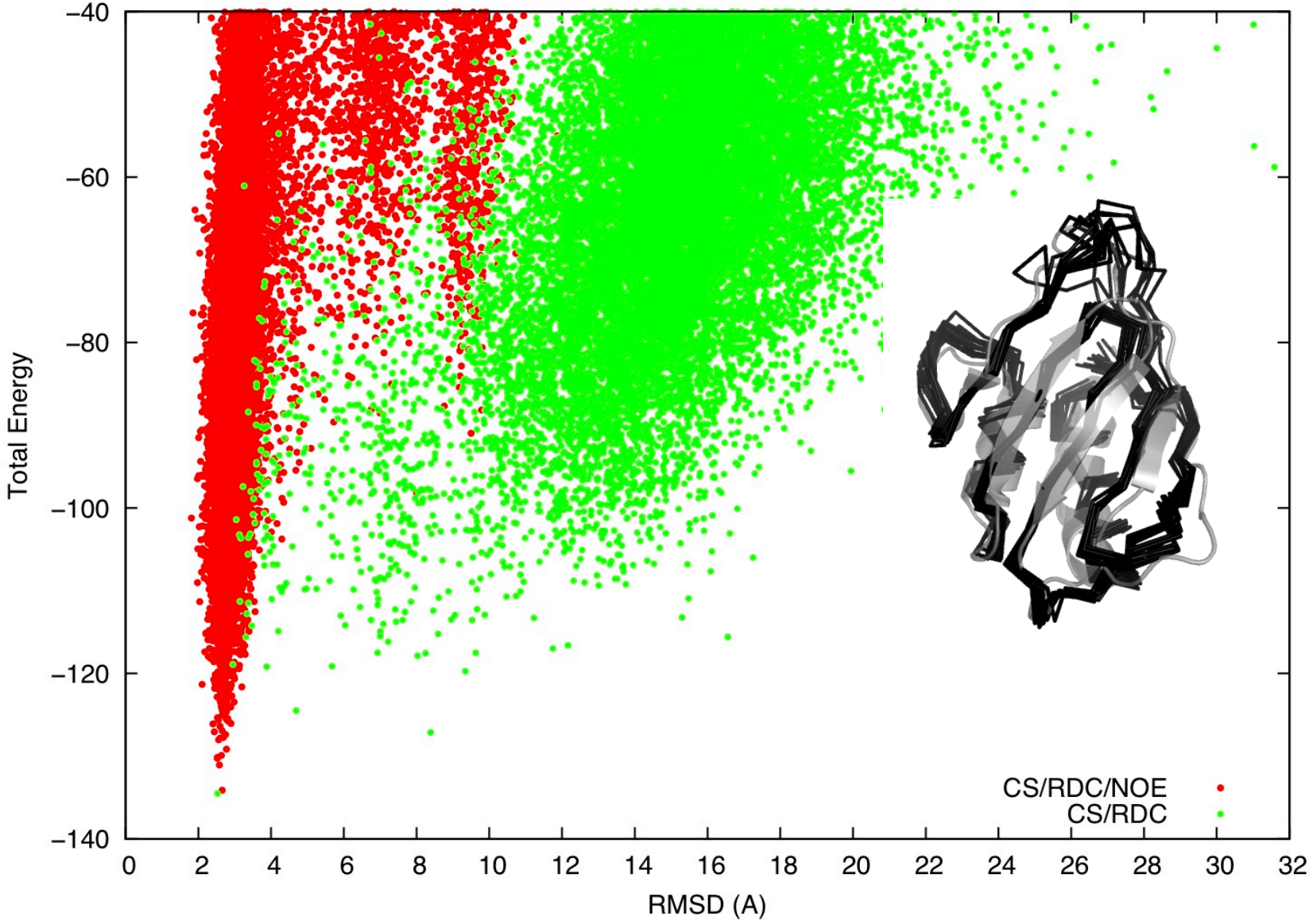
[Saupe matrix file]

```
DATA SAUPE  2.3938e-04  1.0655e-03 -3.4936e-04  4.1407e-04  1.1462e-03
DATA SAUPE  4.2052e-04  1.8428e-03 -5.3311e-04 -3.3843e-04  7.6741e-04
DATA SAUPE  7.5595e-04 -3.7494e-04  6.4788e-04  2.5292e-04 -5.3176e-04
```

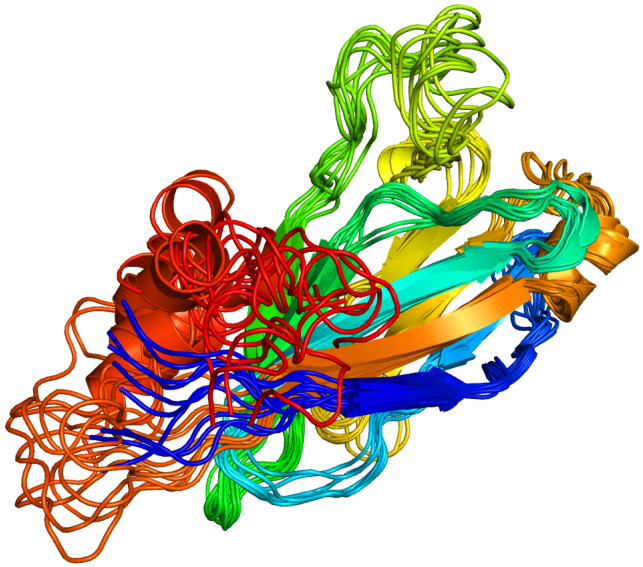
[linear coefficients file]

```
 0.64978552  0.66505875 -0.36807014
-0.34546448 -0.17293961 -0.92235903
 0.67707685 -0.72649071 -0.11738052
```

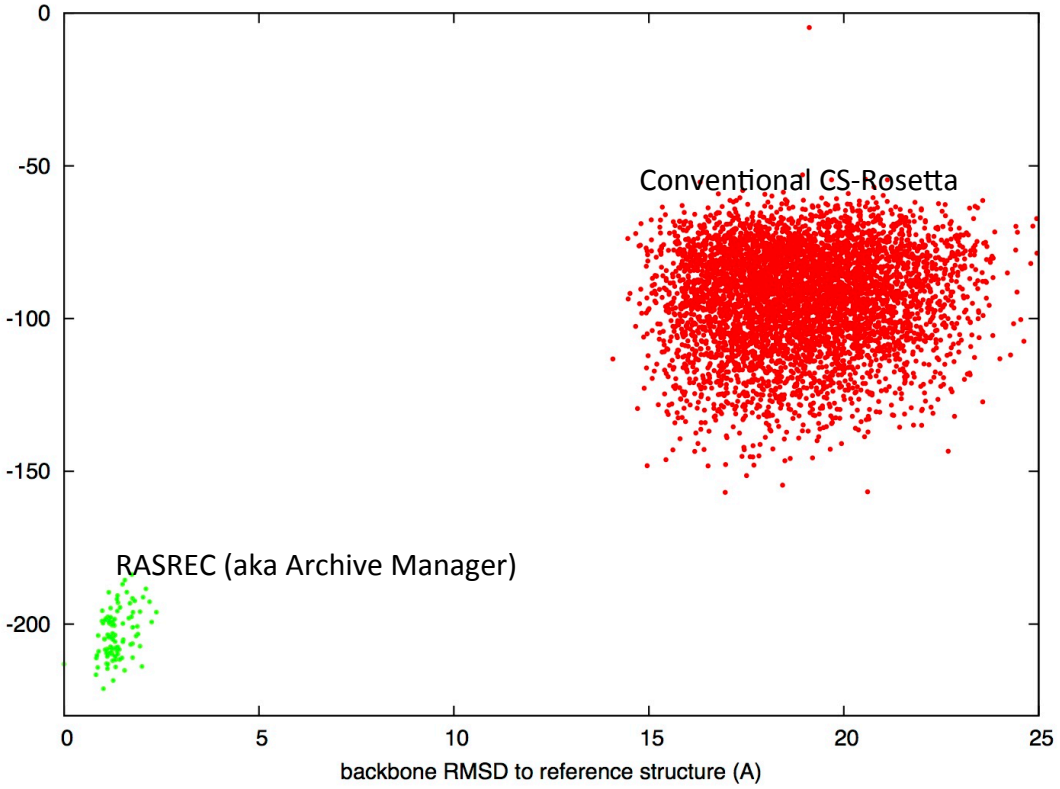
Sparse NOE and RDC data drastically improve sampling of the native state

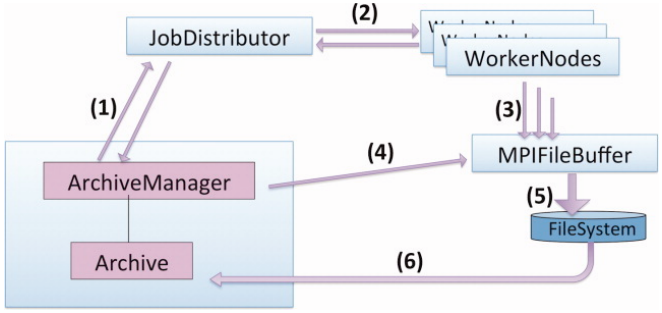


RASREC-Rosetta drastically improves the sampling of native-like features in parallel calculations



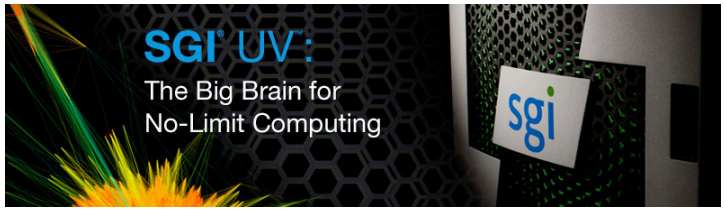
m04/gp34 Immunoevasin





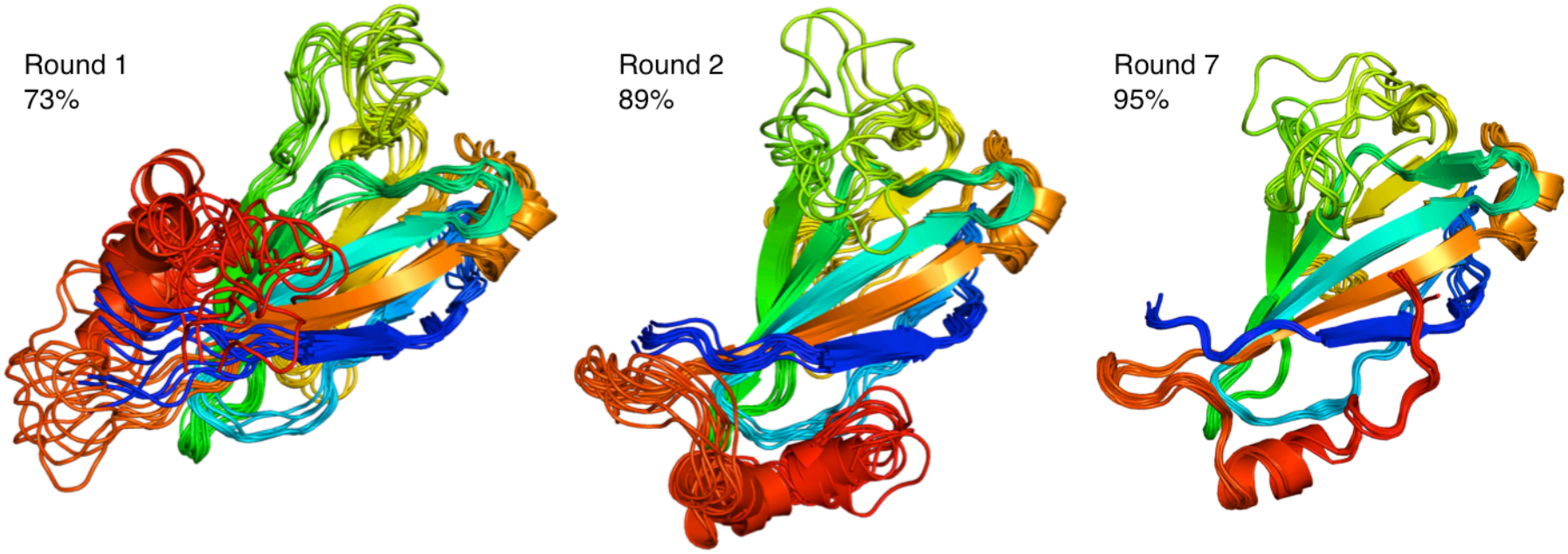
Lange and Baker, *Proteins*, 2012

MPI, 200 cores, 6-8hrs



Oliver Lange

Iterations of Rosetta calculations using different sparse datasets lead to a converged m04ED structure

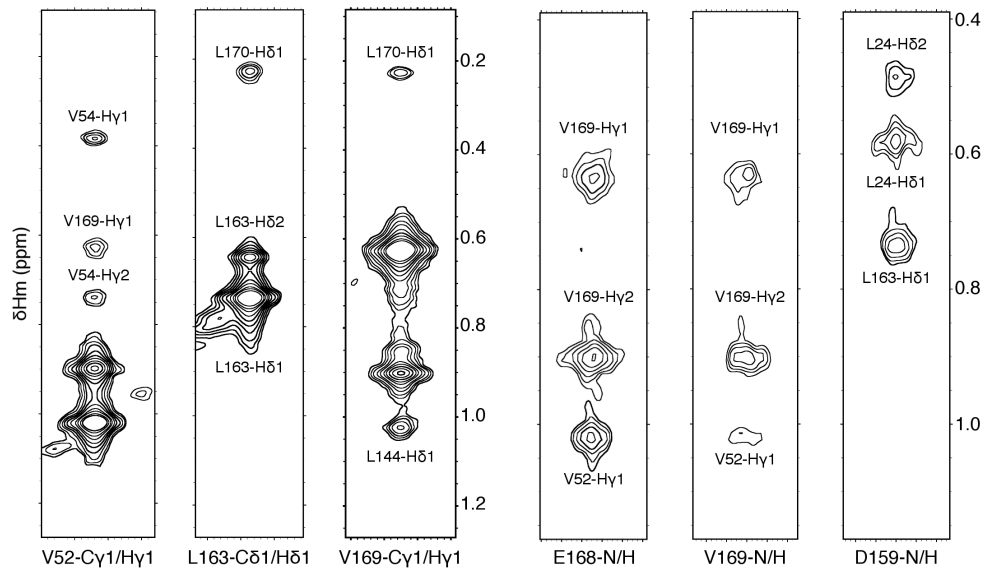


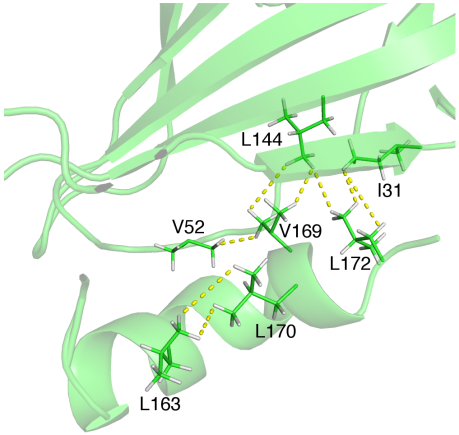
#	Experimental restraints used	Converged residues*	$F_{converged}$	$<E_{Rosetta}>$**	$<Q_{work}>$	$<Q_{free}>$***	Ct-helix^{+}
1	NOE_{amide}	4-74, 82-97, 103-129	73%	-212±4	1.43	1.23	Top^{++}
2	NOE_{amide}, RDC_1	3-74, 85-150	89%	-224±2	0.40	0.68	Under
3	NOE_{amide}, RDC_1, RDC_2	1-75, 83-150	91%	-231±4	0.37/0.41	N/A	Under
4	NOE_{amide}, $NOE_{met\text{-}amide}$, RDC_1	1-74, 83-154	94%	-225±3	0.38	0.58	Under
5	NOE_{amide}, $NOE_{met\text{-}amide}$,$NOE_{met\text{-}met}$, RDC_1	1-74, 83-154	94%	-229±1	0.44	0.60	Under
6	NOE_{amide}, $NOE_{met\text{-}amide}$,$NOE_{met\text{-}met}$, RDC_1, RDC_2	1-74, 83-155	95%	-236±4	0.36/0.40	N/A	Under

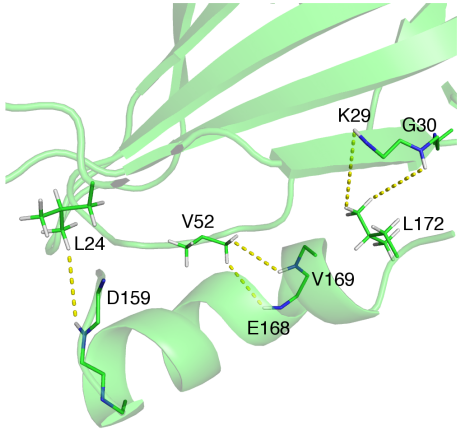
m04 RASREC online tutorial coming soon

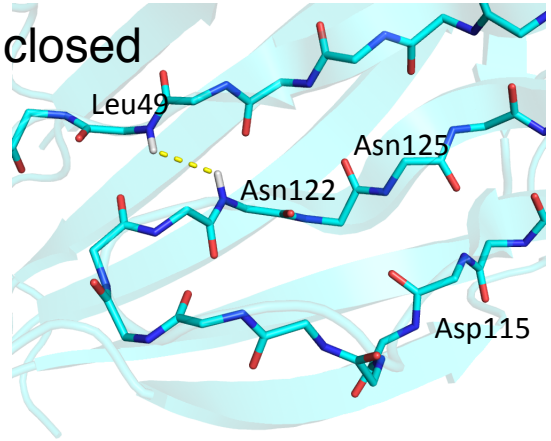
Sgourakis et. al, *Structure, 2015*

NOEs and RDCs provide complementary structural restraints to allow the placement of secondary structure elements

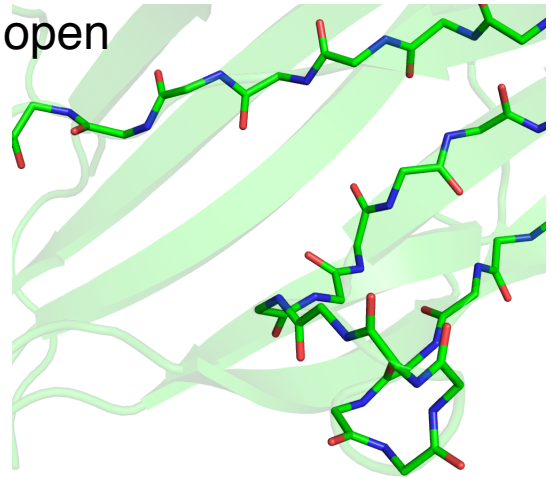




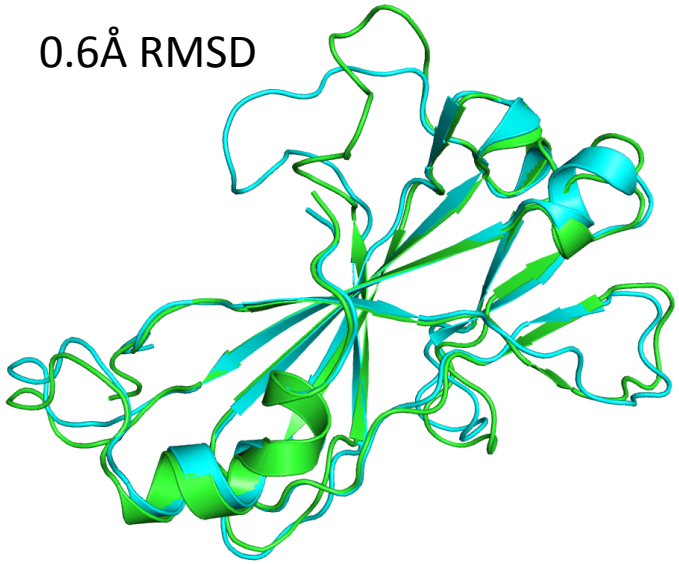




$RMSD^{RDC}_{loop}$: 5.8 / 8.5Hz



$RMSD^{RDC}_{loop}$: 3.6 / 3.7Hz



NMR (Sgourakis et al., *Structure*)
X-ray (Berry et al., *J.Biol.Chem.*)

Structure
Previews

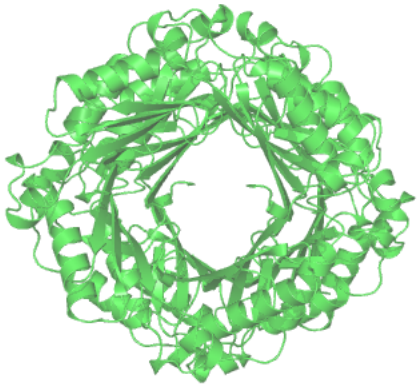
CellPress

Less Is More: Structures of Difficult Targets with Minimal Constraints

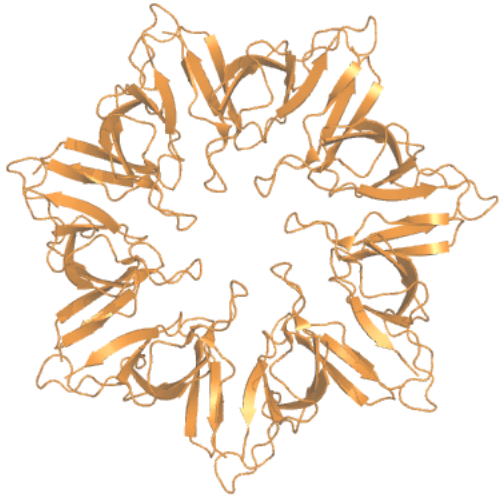
Neil R. Lloyd[1] and Deborah S. Wuttke[1,*]
[1]Department of Chemistry and Biochemistry, University of Colorado, 596 UCB, Boulder, CO 80309, USA
*Correspondence: deborah.wuttke@colorado.edu
http://dx.doi.org/10.1016/j.str.2014.08.004

By merging recent experimental and computational methodology advances, resolution-adapted structural recombination Rosetta has emerged as a powerful strategy for solving the structure of traditionally challenging targets. In this issue of *Structure*, Sgourakis and colleagues solve the structure of one such target, the immunoevasin protein m04, using this approach.

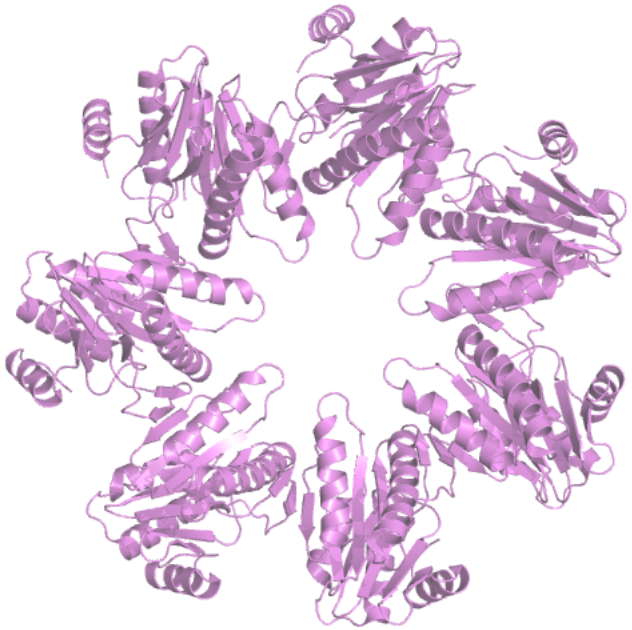
High-MW molecular assemblies are now visible by NMR



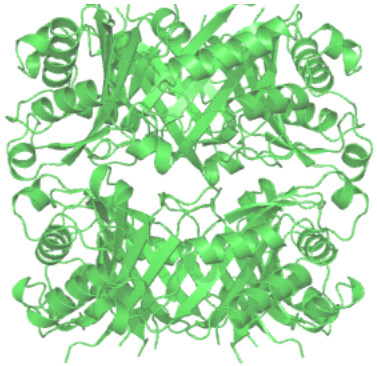
DHNA
(D4, 121 residues)

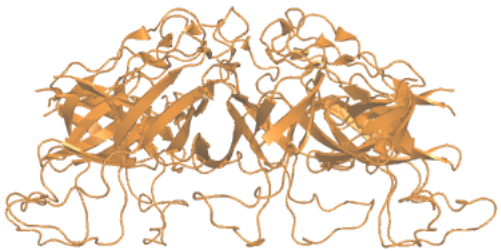


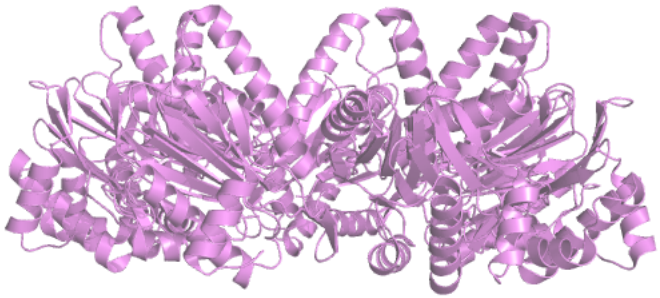
C8, 97 residues



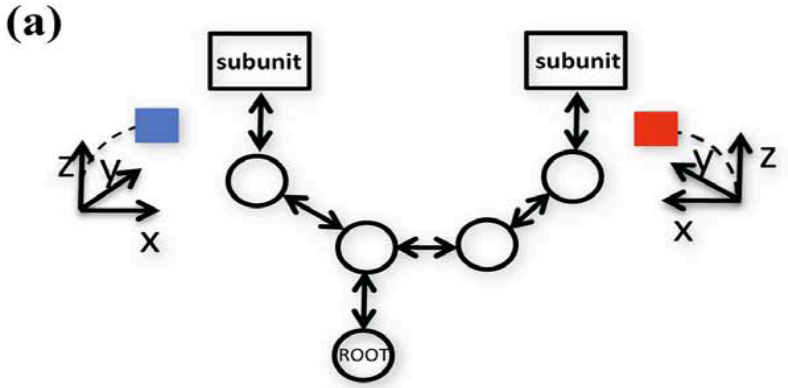
Proteasome
(C7 221 residues)



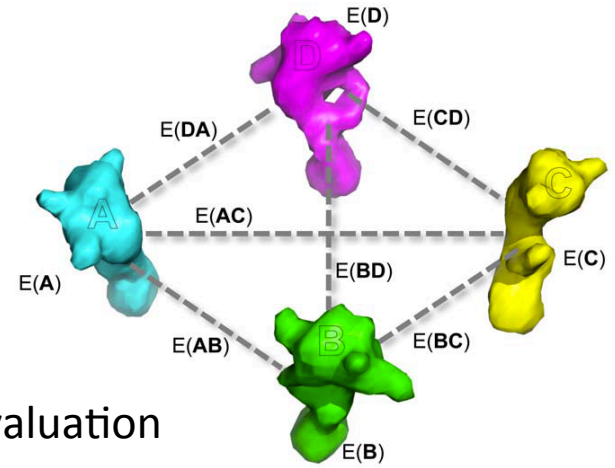




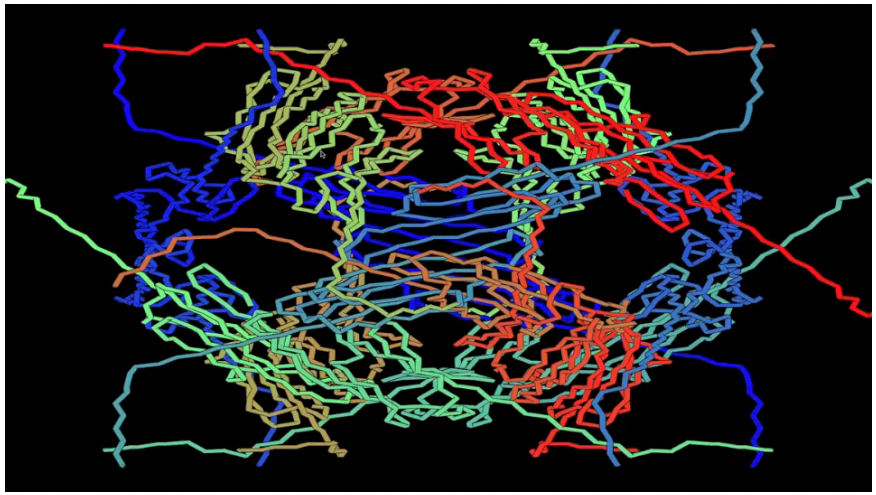
Rosetta3: A powerful symmetric modeling machine



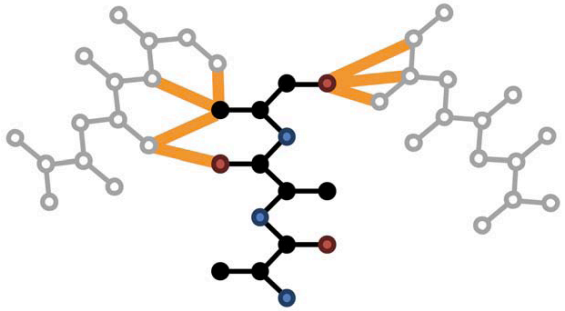
Kinematics



Energy evaluation



Aβ crystalin: 24 subunits, O2 symmetry



Sidechain optimization

Frank DiMaio

DiMaio et al, *PLoS ONE*, 2012

Rosetta Symmetry Basics

Symmetry in Rosetta is encoded in a symmetry definition file (.symm) which is custom-tailored for a given symmetry type (eg. C2, D4, helical, octahedral, etc.).

Briefly, the file contains information for evaluating the energy of the system, the placement of "jumps" connecting the different subunits, and the degrees of freedom to be optimized in the calculation.

Using a symmetry definition file, and the PDB coordinates of a monomeric subunit, Rosetta has all the information require to:

a) reconstruct the entire assembly system and
b) evaluate the energy of the entire system

Scripts that are part of the standard distribution can detect the symmetry information from a pre-exiting PDB file in a way that can then be presented to Rosetta or built given symmetry types *de novo:*

```
perl ~/rosetta/rosetta_source/src/apps/public/symmetry/make_symmdef_file.pl -m helix -a 0 -b N -r 999 -t 14 -p
monomer_001.pdb  >helix29.symm
```

```
~/rosetta/rosetta_source/src/apps/public/symmetry/make_symmdef_file_denovo.py
```

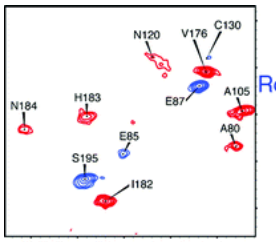
A good practice is to always rebuild the PDB of the entire system starting from monomer coordinates + symdef using the score application:

```
score_jd2.linuxgxxrelease -in:file:s monomer.pdb -symmetry_definition helix29.symm -out:pdb
```

this will reconstruct a file a monomer_0001.pdb containing the coordinates of the entire assembly.

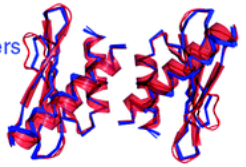
DiMaio et al., *PLOS ONE*, 2011

RosettaOligomers



RDCs, Chemical shifts

RosettaOligomers



Oligomer protein structures

Run CS-Rosetta (monomer)

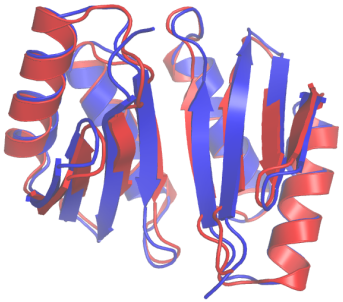
Convergence within 3Å RMSD ?

Yes

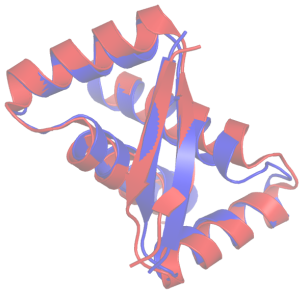
No

Symmetric Docking/ backbone optimization

Fold and Dock using RDCs



Convergence within a single cluster ?

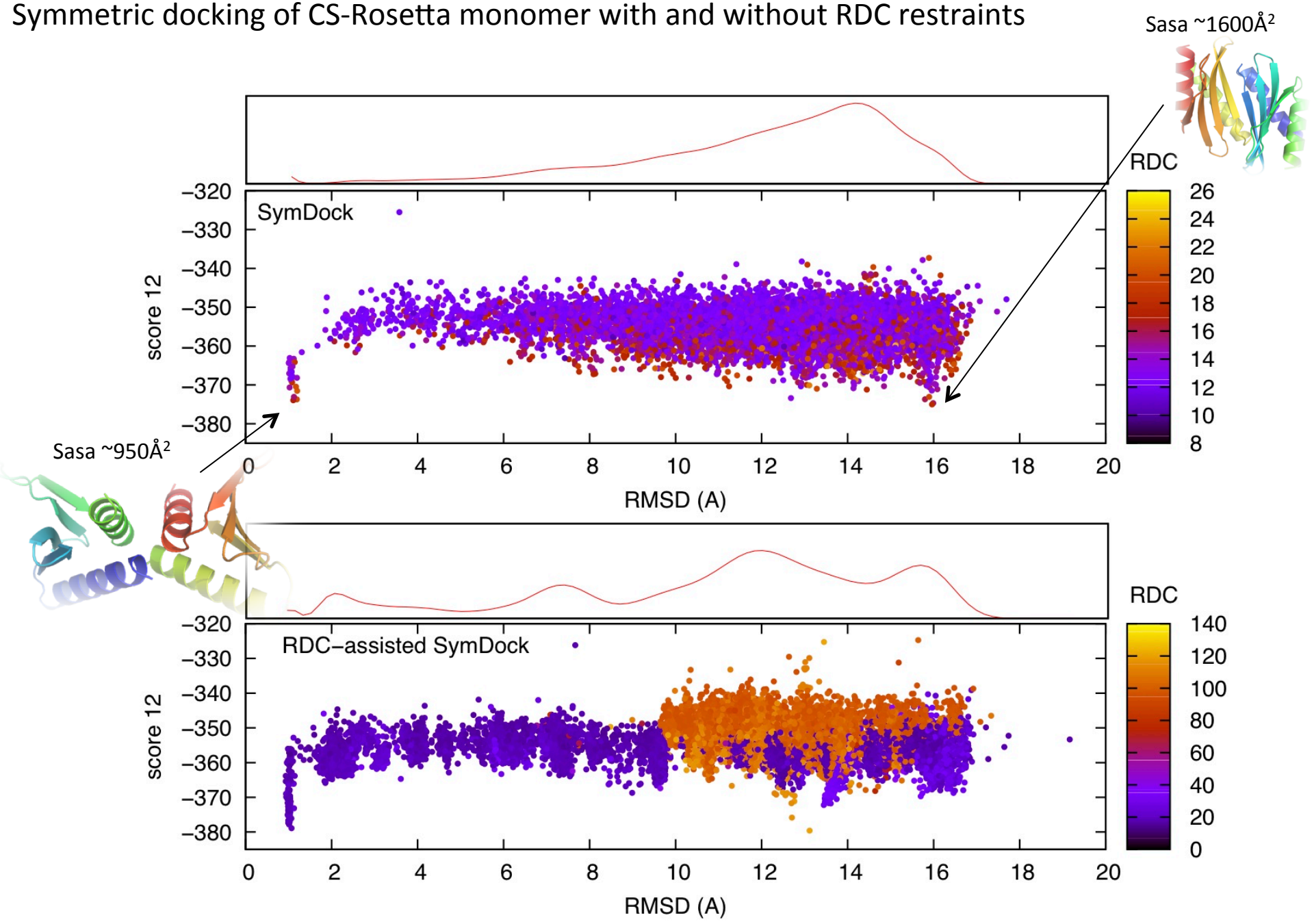


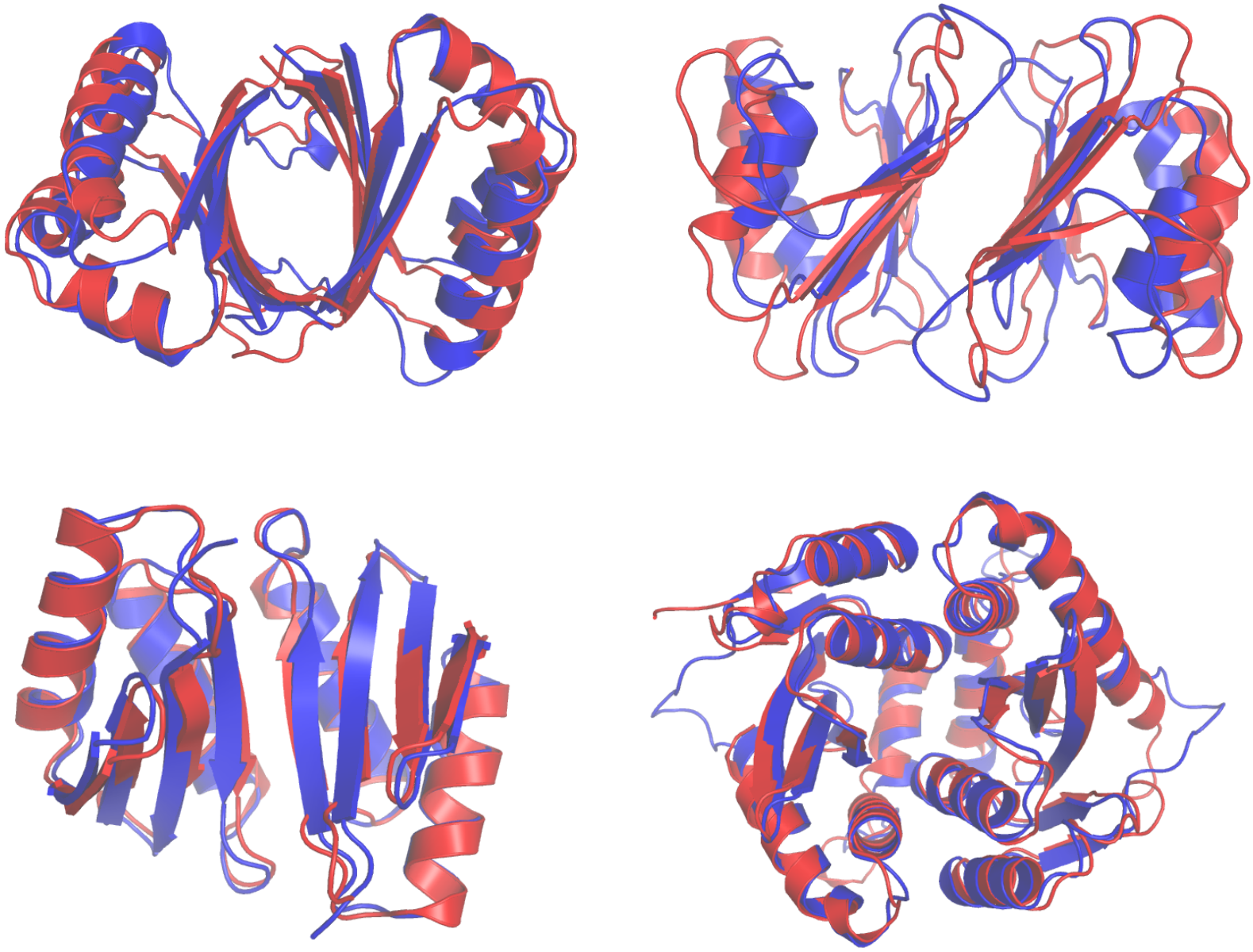
Yes

Final ensemble

Sgourakis et al., *JACS, 2010*

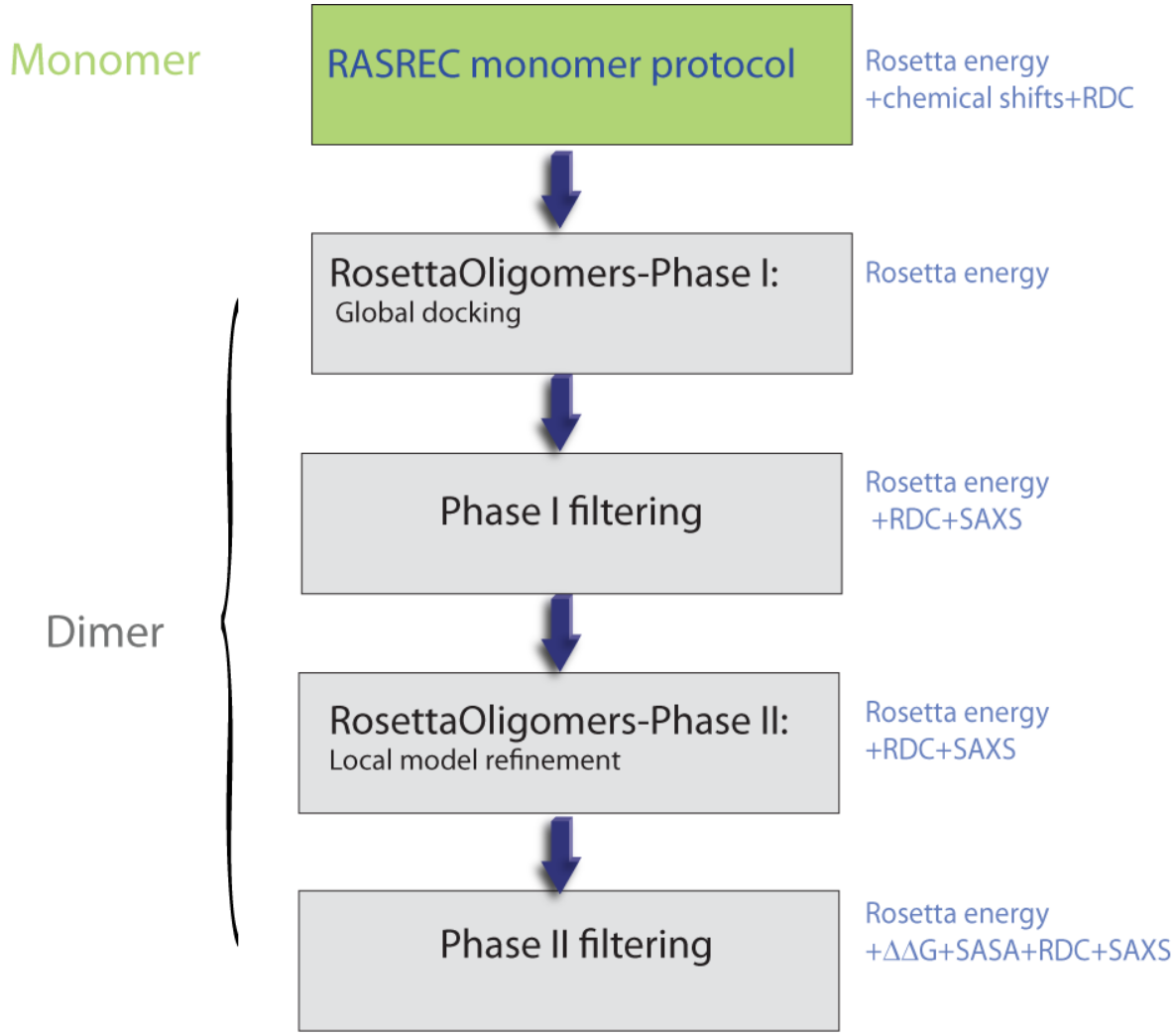
Symmetric docking of CS-Rosetta monomer with and without RDC restraints





Sgourakis et al., *JACS, 2011*

Combining RDCs and SAXS with the Rosetta Energy function and advanced sampling methods



Rossi, Shi, .., Sgourakis, *Proteins, 2015*

A SAXS score can be turned on/off during the calculations

Global sampling mode (Phase I):

-run:protocol symdock

-in:file:s *monomer_input_rasrec.pdb*

-symmetry:symmetry_definition C2.sym

-symmetry:*initialize_rigid_body_dofs*

FILTER BY SAXS, RDC

Local Perturbation mode (Phase II):

-run:protocol symdock

-in:file:s sel_monomer_from_phase1.pdb

-symmetry:symmetry_definition C2.sym

-symmetry:*perturb_rigid_body_dofs 3 5*

-docking:low_patch *[rdc+saxs]*

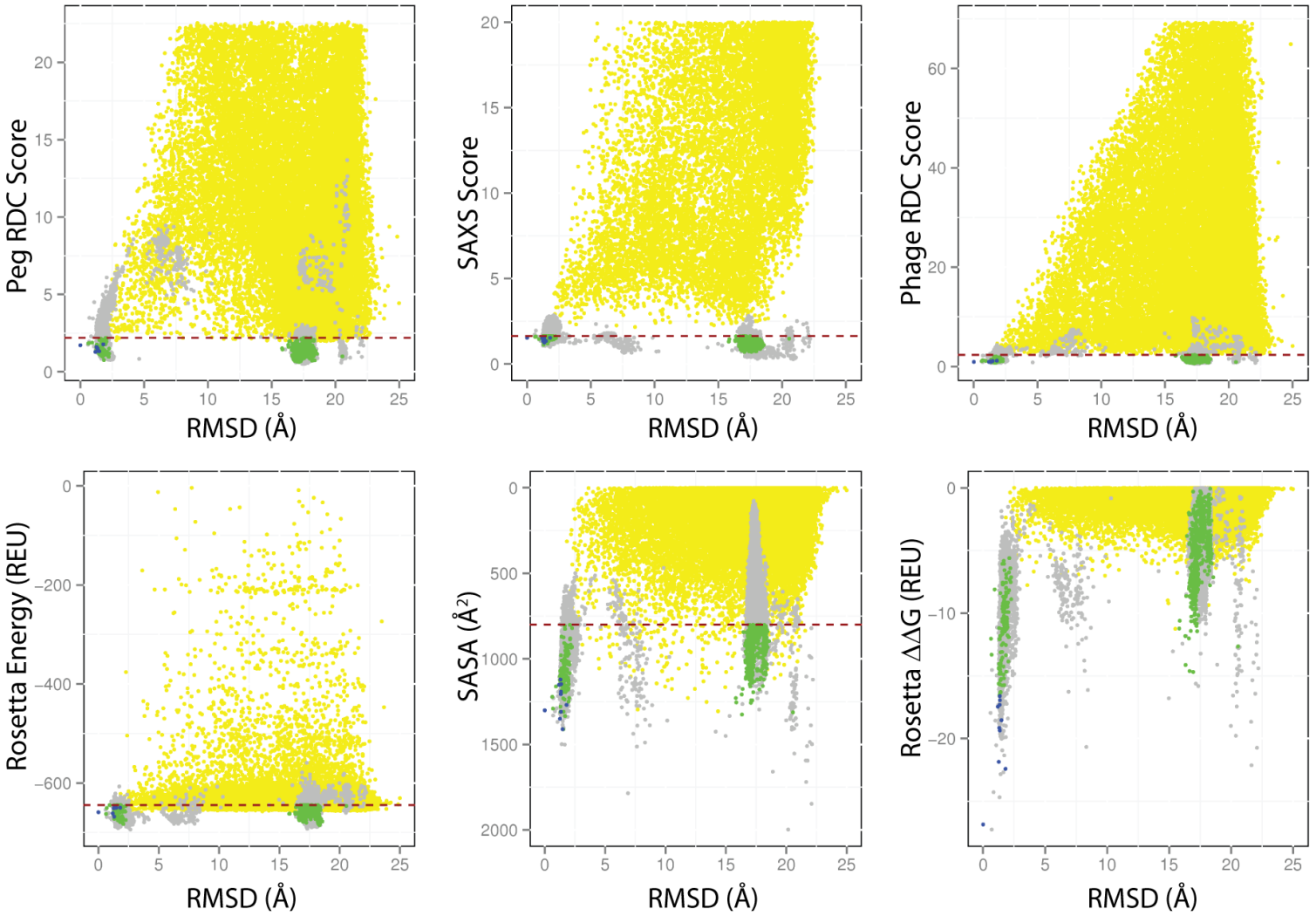
-docking:high_patch *[rdc+saxs]*

-docking:high_min_patch *[rdc]*

-docking:pack_patch *[rdc]*

Rossi, Shi, .., Sgourakis, *Proteins, 2015*

Conformational coverage in two-tier docking protocol



Phase I: global docking
Phase II: local perturbation/refinement
Phase II post-filter
Final Ensemble of Models

$$E_{Tot} = E_{Rosetta} + w_{EXP} * (w_1 E_{RDC1} + w_2 E_{RDC2} + w_3 E_{SAXS})$$

Implementation of SAXS restrains

Uses a coarse-grain representation with residue-specific “form factors” (Stovgaard et al., BMC Bioinformatics, 2010):

-residues:patch_selectors CENTROID_HA

-score:saxs:ref_spectrum saxs_sparse.dat

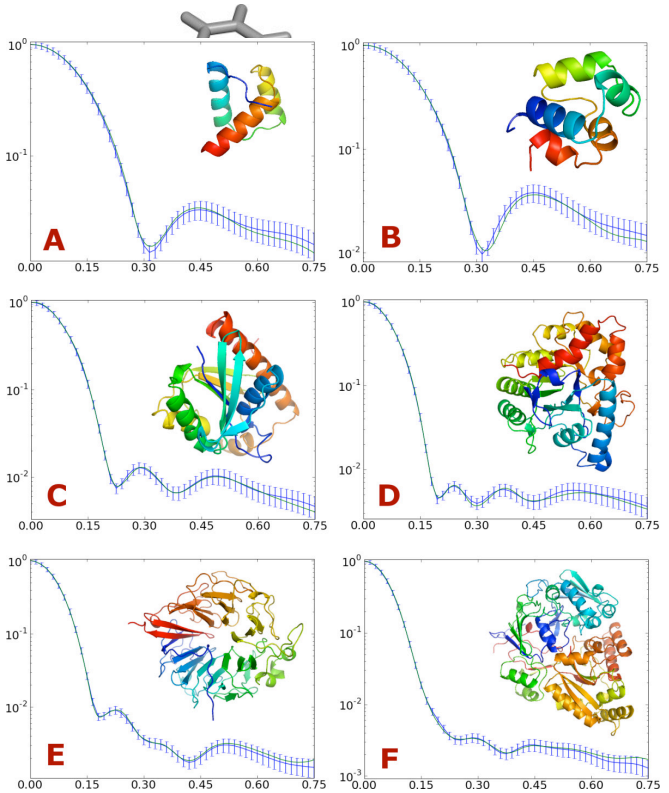
-score::patch patch_saxs

Patch file contains:

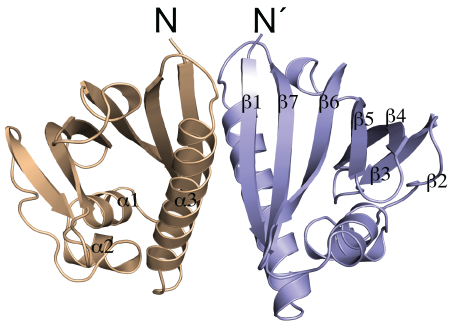
fastsaxs = 0.05

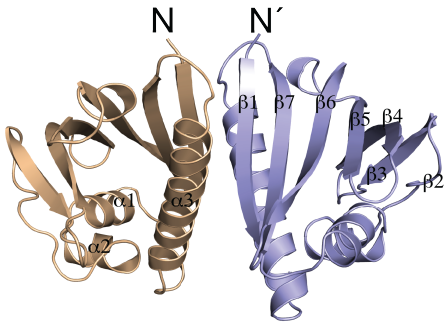
Data file:

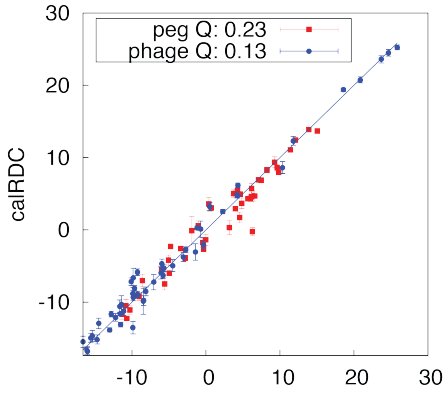
```
# q I(q) Delta
0.00771096 7554.24 70.6635
0.017006 7253.15 9.33698
0.0263011 6830.58 7.28595
0.0355961 6285.76 6.17379
0.0448912 5670.72 5.27
0.0541862 4985.43 4.56433
0.0634813 4285.38 3.83601
0.0727763 3587.02 3.27892
0.0820714 2926.51 2.97787
```

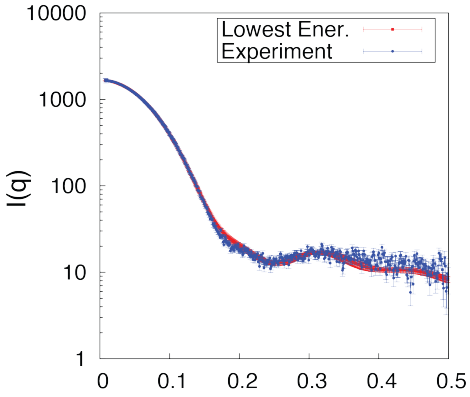


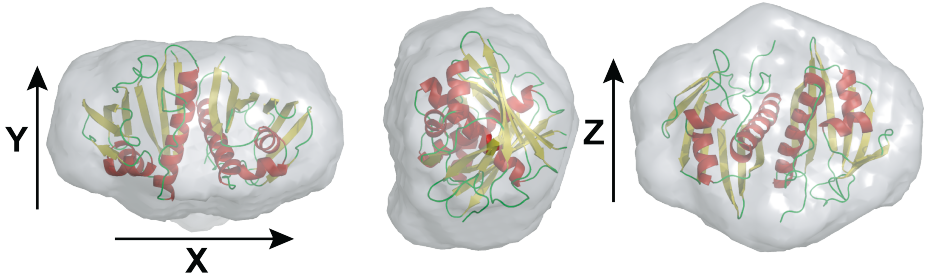
Model fits


N
N´
β1
β7
β6
β5
β4
β3
β2
α1
α3
α2


N
N´
β1
β7
β6
β5
β4
β3
β2
α1
α3
α2


peg Q: 0.23
phage Q: 0.13
calRDC
30
20
10
0
-10
-10
0
10
20
30


Lowest Ener.
Experiment
I(q)
10000
1000
100
10
1
0
0.1
0.2
0.3
0.4
0.5


Y
X
Z

Coming up…

Python interface to setup and run from command line:

CS/HM-Rosetta

*RASREC Rosetta**

RosettaOligomers

Online documentation

Sparse ILV+RDC data tutorial

**requires MPI-enabled computer cluster*

sgourakisn@mail.nih.gov
www.csrosetta.org

Acknowledgements

Oliver Lange, Frank DiMaio, James Thompson, Lei Shi, Yang Shen, Robert Vernon

Paolo Rossi, Guy Montelione (Rutgers)
Antoine Loquet, Jean-Philippe Demers, Birgit Habenstein,
Karin Giler, Suresh Kumar-Vasa, Stefan Becker**,** Adam Lange (MPI)
Kannan Natarajan, David Margulies (NIAID, NIH)
Justin Benesch, Andy Baldwin (Oxford)
Julien Bergeron, Liam Worrall, Natalie Strynadka (UBC)
Omry Morag, Amir Goldbourt (Tel Aviv University)
Wei Qiang (SUNNY Binghampton)
Nick Fitzkee (Missisipi State)

David Baker, Ad Bax

NIDDK and NIAID Intramural Research Programs
NIAID Career Transition Award (K22) AI 112573-01

NIH
National Institute of Diabetes and Digestive and Kidney Diseases

UNIVERSITY OF WASHINGTON

National Institute of Allergy and Infectious Diseases