

Protein Folding

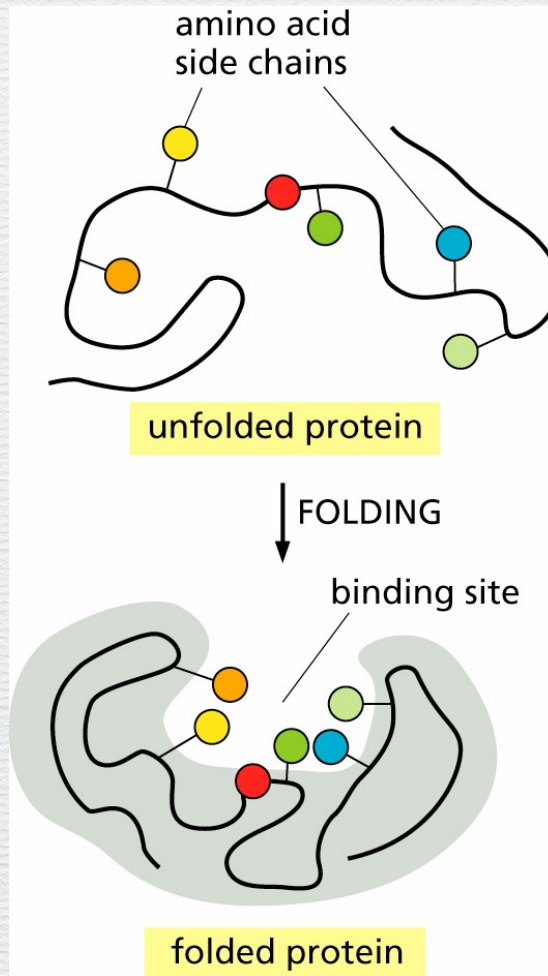
To read:

● http://en.wikipedia.org/wiki/Protein_folding

- Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science. 2005 Sep 16;309(5742):1868-71
- Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PLoS One. 2011;6(12):e28766.

Slides from Howard Feldman hfeldman@blueprint.org

Protein folding



Folding movie

THEORETICAL AND COMPUTATIONAL BIOPHYSICS GROUP

NIH Center for Macromolecular Modeling and Bioinformatics

www.ks.uiuc.edu

presents

Six Microseconds of Protein Folding

http://www.youtube.com/watch?annotation_id=annotation_957349&feature=iv&src_vid=AlfvWESPyZY&v=sD6vyfTtE4U

Ab Initio Protein Structure Prediction

- Predicting the 3D structure of a protein without any “prior knowledge”
- Uses when homology modeling not is possible.
- Equivalent to solving the “Protein Folding Problem”
- Similar methods useful for “Protein design”
 - Protein design is the “inverse” protein folding problem, i.e design a sequence that fold into a given fold.
 - Potentially easier and more useful

ab-initio protein structure prediction

■ Optimization problem

- Define some initial model.
- Define a function mapping structures to numerical values (the lower the better).
- Solve the computational problem of finding the global minimum.

■ Simulation of the actual folding process

- Build an accurate initial model (including energy and forces).
- Accurately simulate the dynamics of the system.
- The native structure will emerge.
- No hope due to large search space

Ab Initio Prediction

- Purists will argue must use laws of physics alone
 - But on what level ?
- However most successful methods use a blend of physics, fold recognition, and statistical probability
- Still an ongoing research problem, but becoming less essential as databases grow
 - But also useful for mini-domains and loop

Ab Initio Folding

- Two Central Problems

- Sampling vast conformational space
- The energy minimum problem

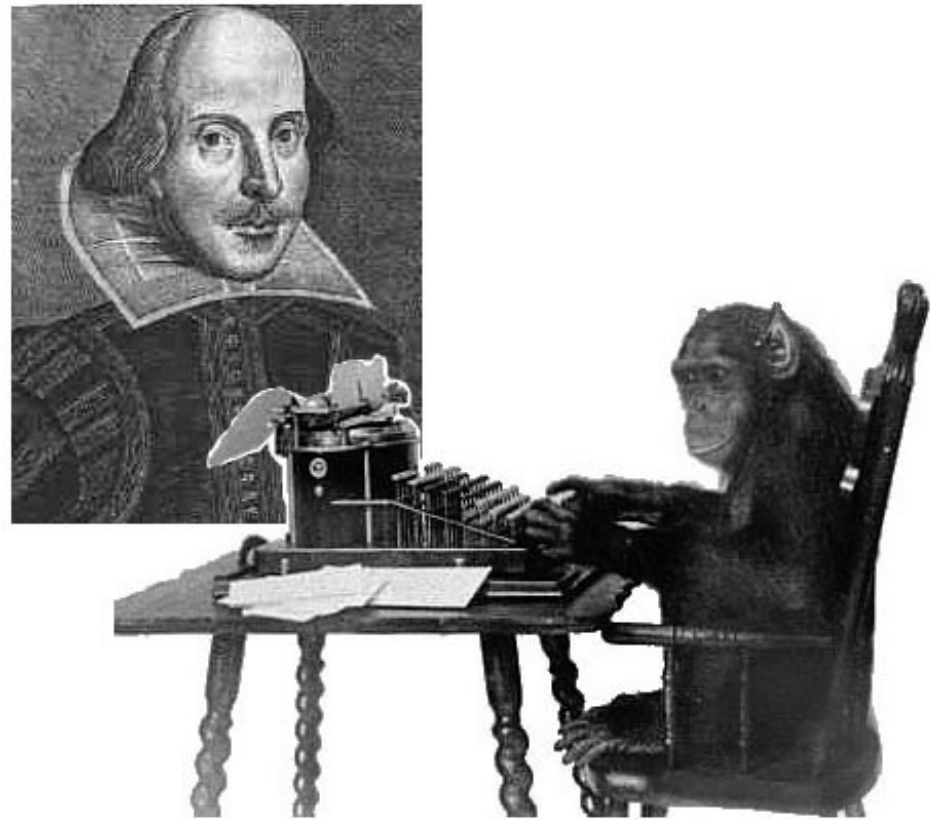
- The Sampling Problem (Solutions)

- Lattice models, off-lattice models, simplified chain methods – exhaustive sampling not possible, even for small peptides

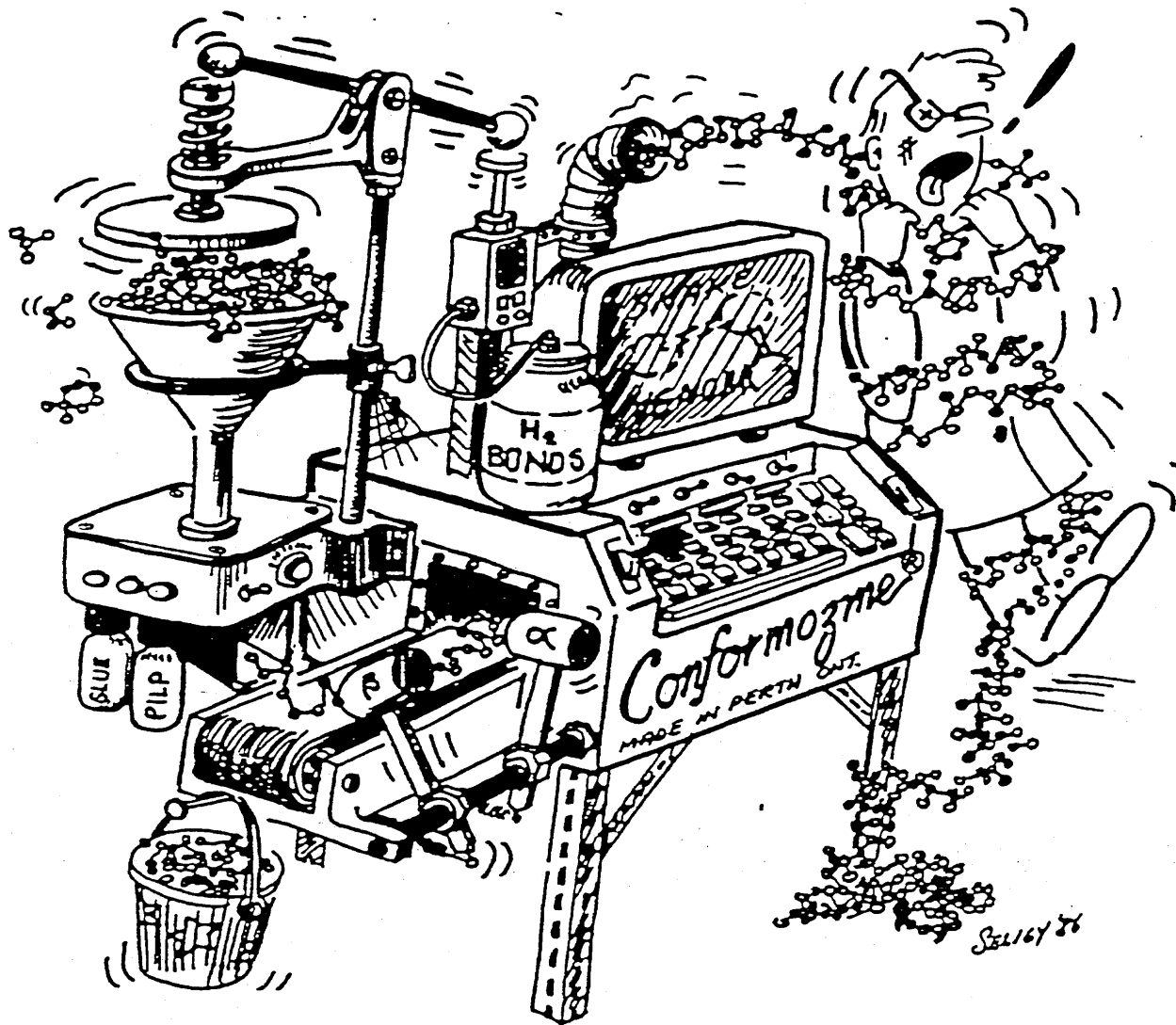
- The Energy Problem (Solutions)

- Threading energies, packing assessment, topology assessment, physics

An infinite
number of
monkeys on
an infinite
number of
typewriters
would eventually



recreate all the works of Shakespeare, and similarly, an infinite number of CPUs could eventually fold every known protein.



CANADA'S FIRST PROTEIN FOLDING
MACHINE!

Molecular mechanics based models

- Could we just use MD simulations to fold proteins.
 - Folding is in the mS to S scale
 - Current simulations is in the μ S scale
 - How accurate are the energy functions
- Folding@home
 - Parallel simulations on distributed computers
 - Many mS of simulations
 - Runs on PS3 (Check our kitchen)
 - Folds small proteins
 - Can not (yet) fold big proteins.
 - Often uses implicit water models

Energy Minimization (Theory)


- Treat Protein molecule as a set of balls (with mass) connected by rigid rods and springs
- Rods and springs have empirically determined force constants
- Allows one to treat atomic-scale motions in proteins as classical physics problems

Folding@Home Distributed Computing - Microsoft Internet Explorer

File Edit View Favorites Tools Help


Back Forward Stop Reload Home Search Favorites

Address <http://folding.stanford.edu/> Links



Folding@home

distributed computing



[Chinese](#) (中文) [Dutch](#) (Nederlands) [French](#) (Français) [German](#) (Deutsch)
[Italian](#) (Italiano) [Japanese](#) (日本語) [Korean](#) (한국말) [Persian](#) (فارسی)
[Portuguese](#) (Português) [Russian](#) (Русский) [Spanish](#) (Español) [Vietnamese](#) (Tiếng Việt)

[Home](#)
[Download](#)
[FAQ](#)
[Forum](#)
[Help!](#)
[Education](#)
[News](#)
[Stats](#)
[Science](#)
[Results](#)

Our goal: to understand protein folding, protein aggregation, and related diseases

What are proteins and why do they "fold"? Proteins are biology's workhorses -- its "**nanomachines**." Before proteins can carry out their biochemical function, they remarkably assemble themselves, or "**fold**." The process of protein folding, while critical and fundamental to virtually all of biology, remains a mystery. Moreover, perhaps not surprisingly, when proteins do not fold correctly (i.e. "misfold"), there can be serious effects, including many well known **diseases**, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, Huntington's, Parkinson's disease, and many cancers and cancer-related syndromes.



Results from Folding@Home

Internet

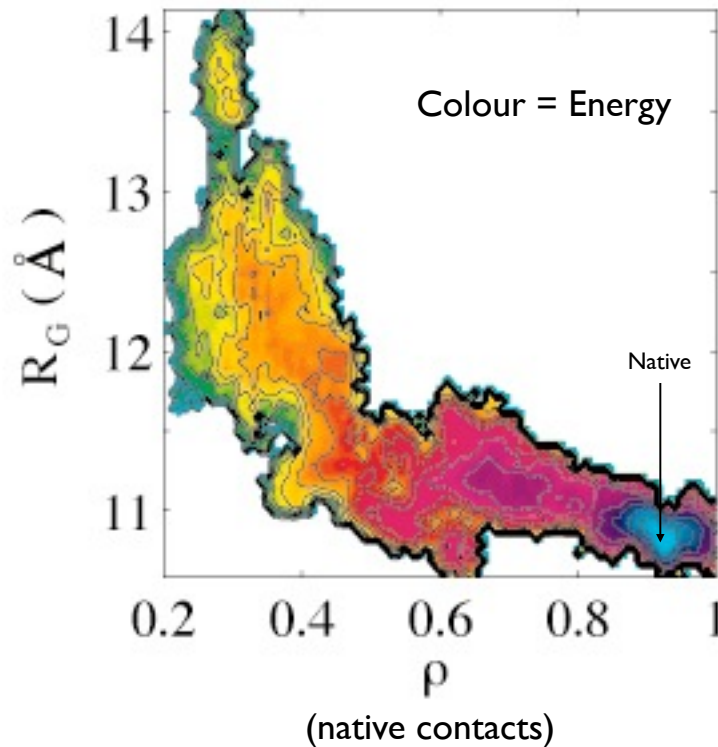
Folding@home Intro

- The work unit uses the cpu to “fold” the protein in millions of combinations and send the results back to Stanford.
- The program then downloads another work unit and repeats.
- On average 1 work unit will take anywhere from a few hours to a few days to complete on a P4 2.6Ghz CPU.

What does Folding@home do?

- Folding@home is a distributed computing project which studies protein folding, misfolding, aggregation and related diseases.
- Folding@home (F@H) uses spare cpu cycles to fold proteins in the form of Work Units (WU) and send the results to Stanford Universities servers.

The L shape of a protein folding pathway



Brooks and
Sheinerman's
Protein G

MD folding
512 Processors
Cray T3E 1 month

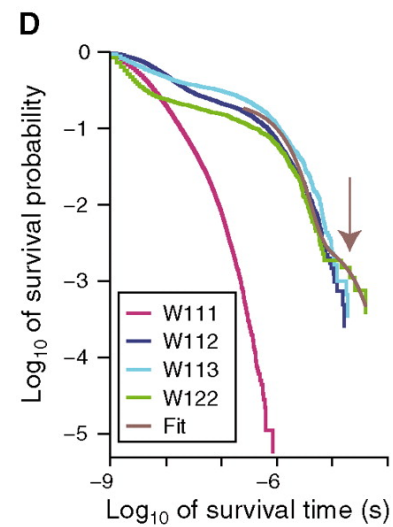
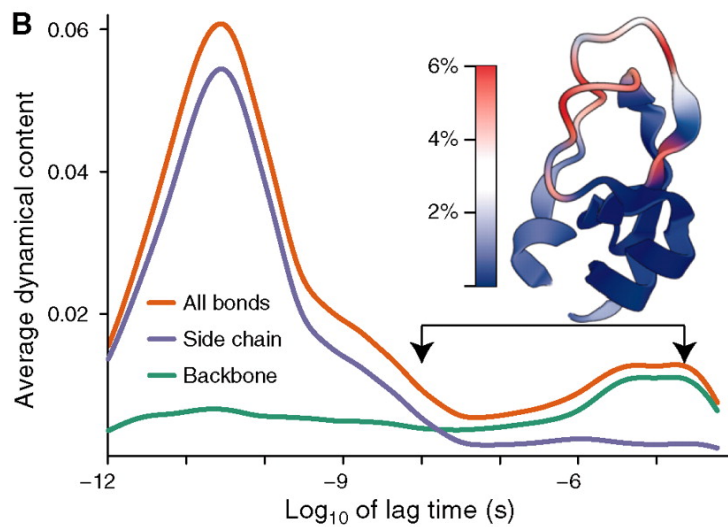
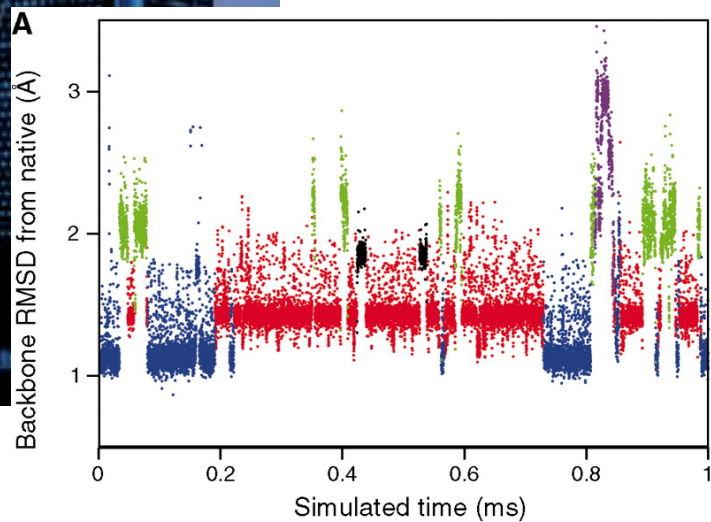
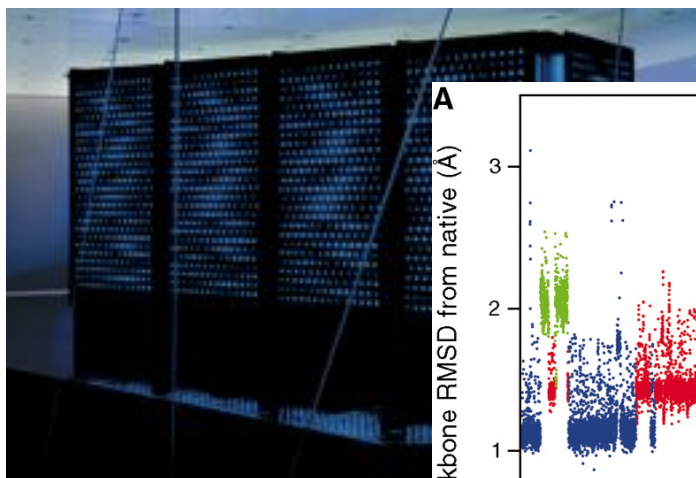
No “core nucleation”
apparent

Folding@home video



<http://www.youtube.com/watch?v=EZlXuOgknuE>

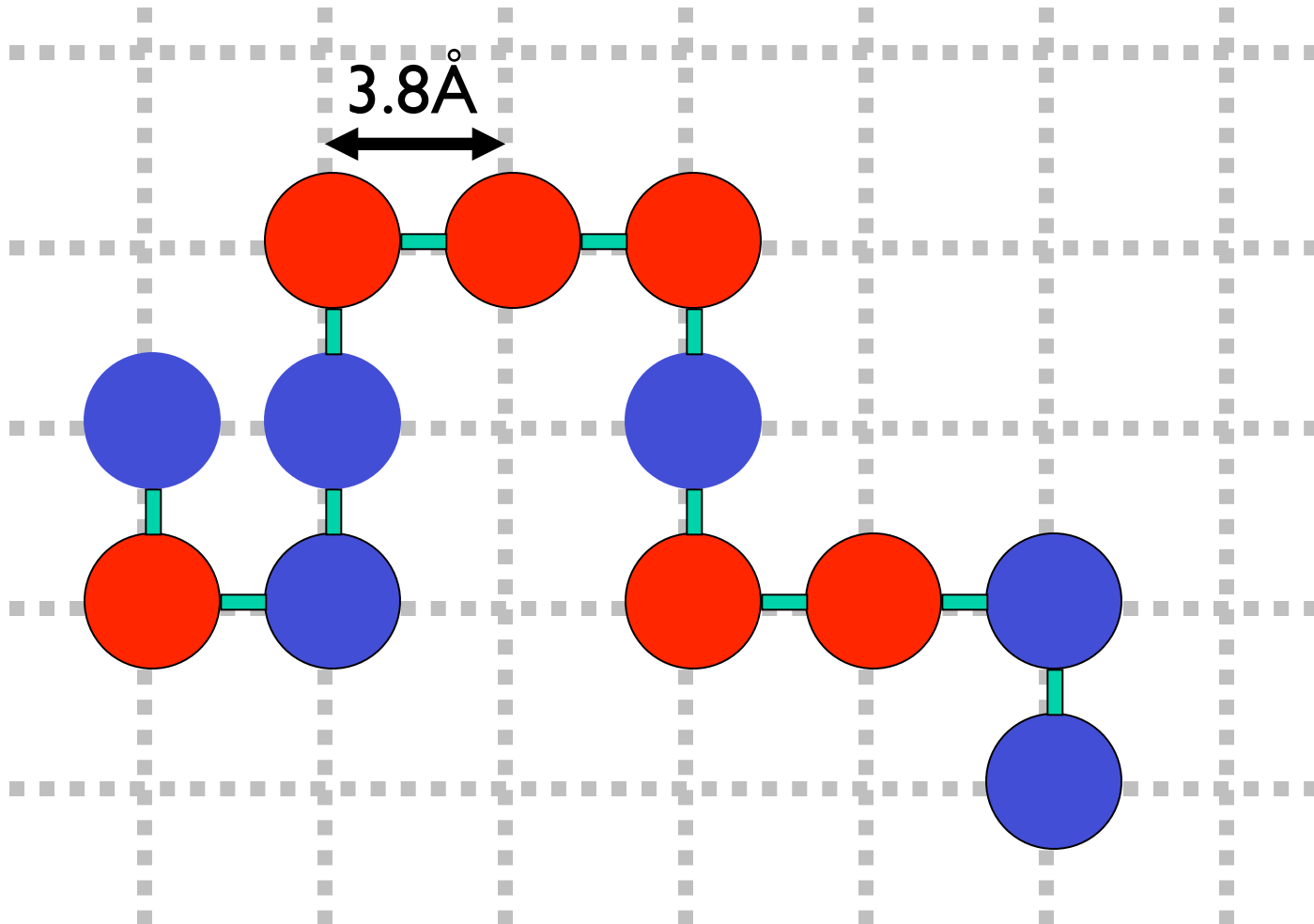
Anton (D.E. Shaw)



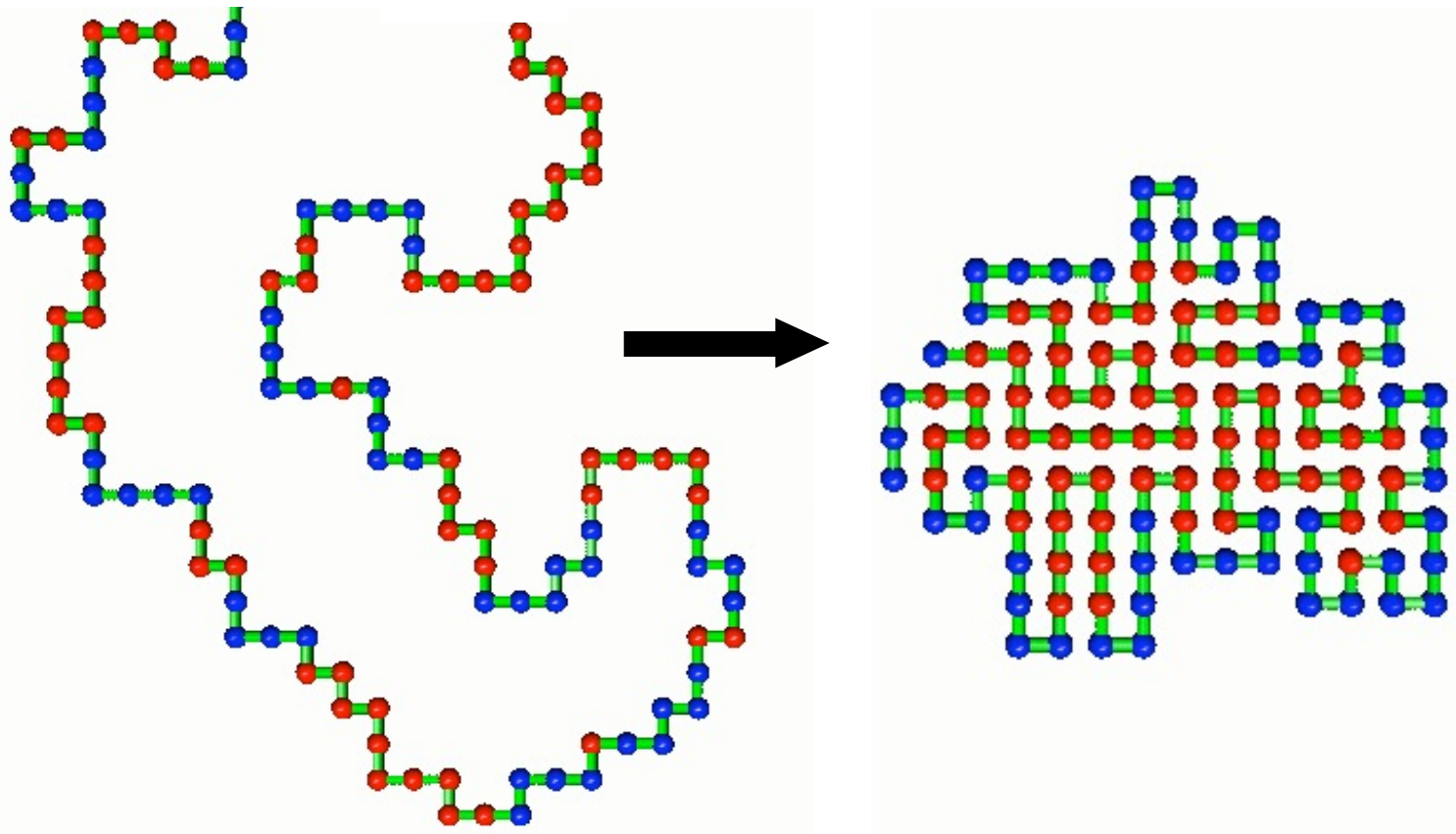
Simplified models

- Can we use simplified models ?
 - How much can we simplify ?
 - Are these describing the physics correctly.
 - What is physically correct
- Lattice models
- Simplified off-lattice-models
- All atom models

A Simple 2D Lattice



Lattice Folding



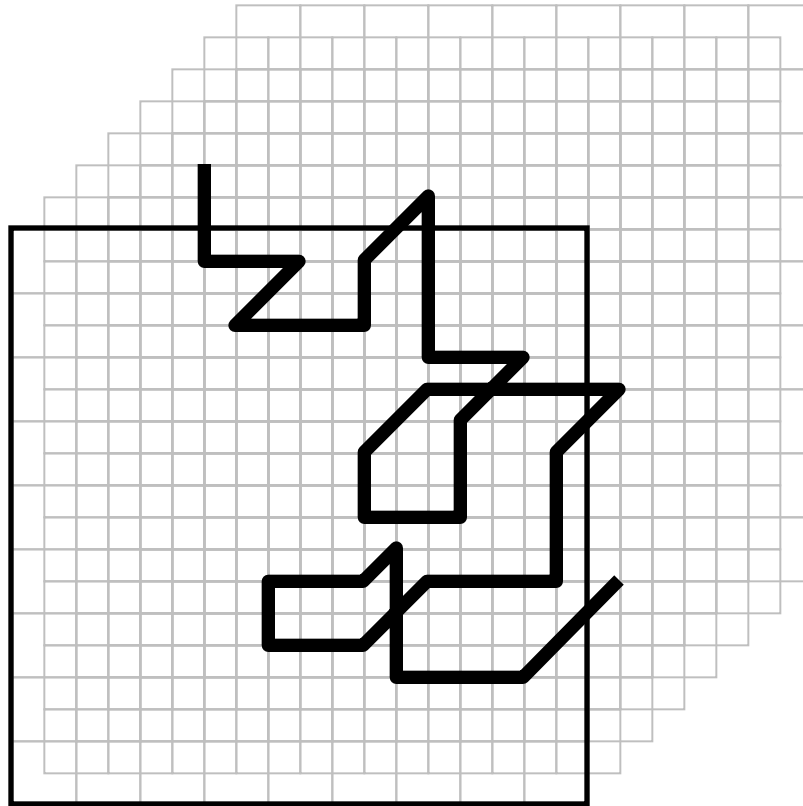
Lattice Algorithm

1. *Build a “ $n \times m$ ” matrix (a 2D array)*
2. *Choose an arbitrary point as your N terminal residue*
3. *Add or subtract “1” from the x or y position of the start residue*
4. *Check to see if the new point (residue) is off the lattice or is already occupied*
5. *Evaluate the energy*
6. *Go to step 3. and repeat until done*

Lattice Energy Algorithm

- *Red = hydrophobic, Blue = hydrophilic*
- *If Red is near empty space $E = E+1$*
- *If Blue is near empty space $E = E-1$*
- *If Red is near another Red $E = E-1$*
- *If Blue is near another Blue $E = E+0$*
- *If Blue is near Red $E = E+0$*

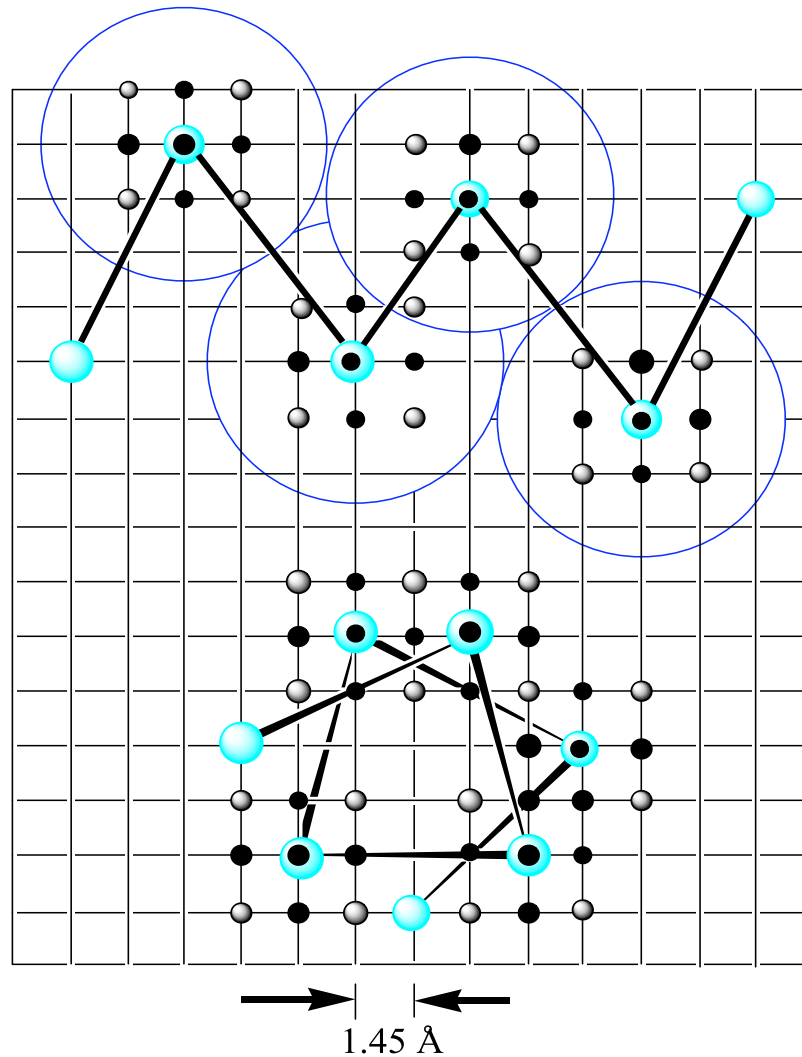
3D Lattices



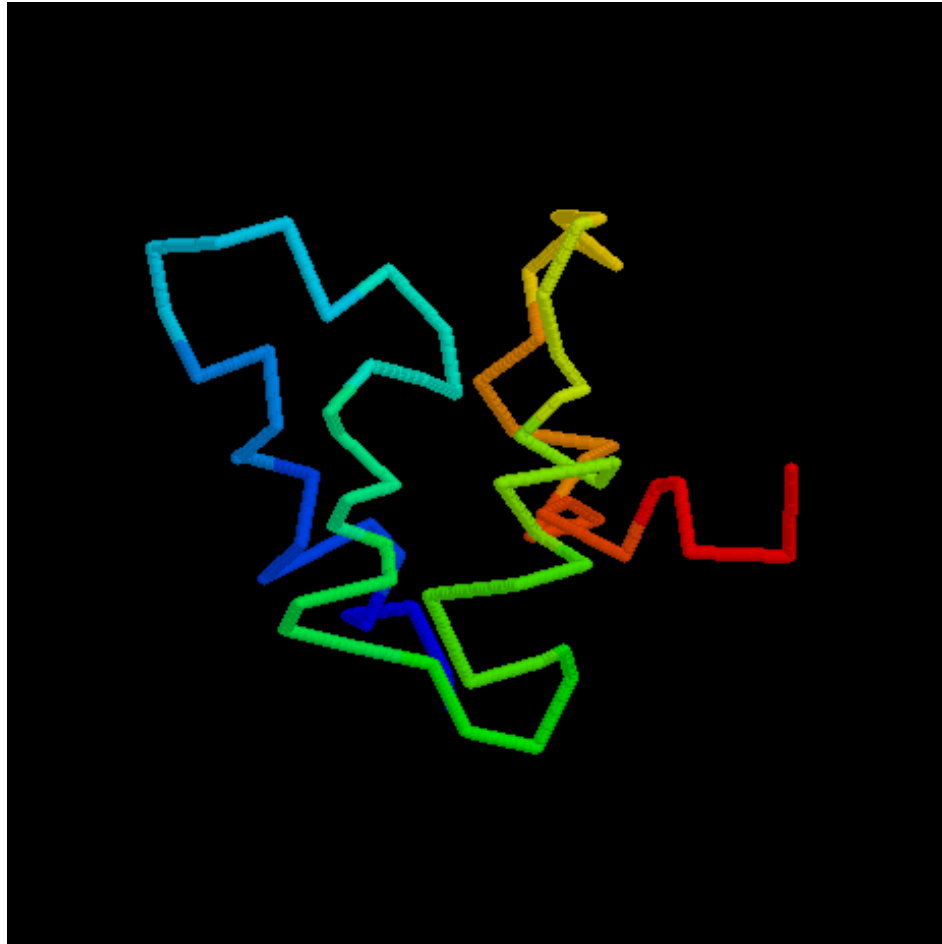
Use of cubic lattice models

- For simple 2D and 3D lattices it is possible to
 - Calculate all possible (compact) structure
 - Design many different sequences
- Study the relationship between sequence and structure
 - What fraction of all sequences fold ?
 - What differentiate folding and non-folding sequences ?
 - What is the “solution” to Levinthals paradox.
- Not really for structure prediction

More Complex Lattices



Really Complex 3D Lattices



J. Skolnick

Lattice Methods

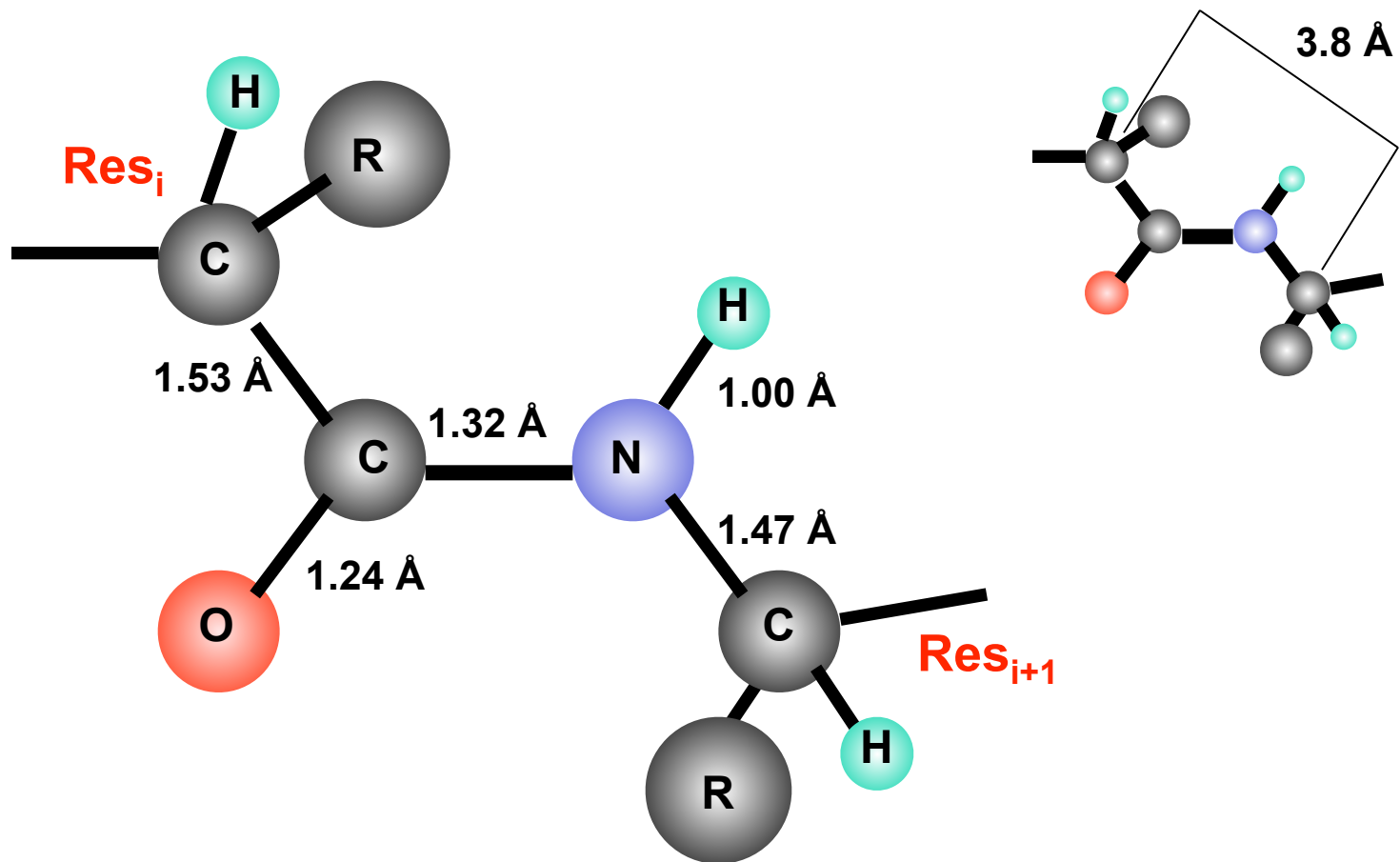
Advantages

- Easiest and quickest way to build a polypeptide
- Implicitly accounts for excluded volume
- More complex lattices allow reasonably accurate representation

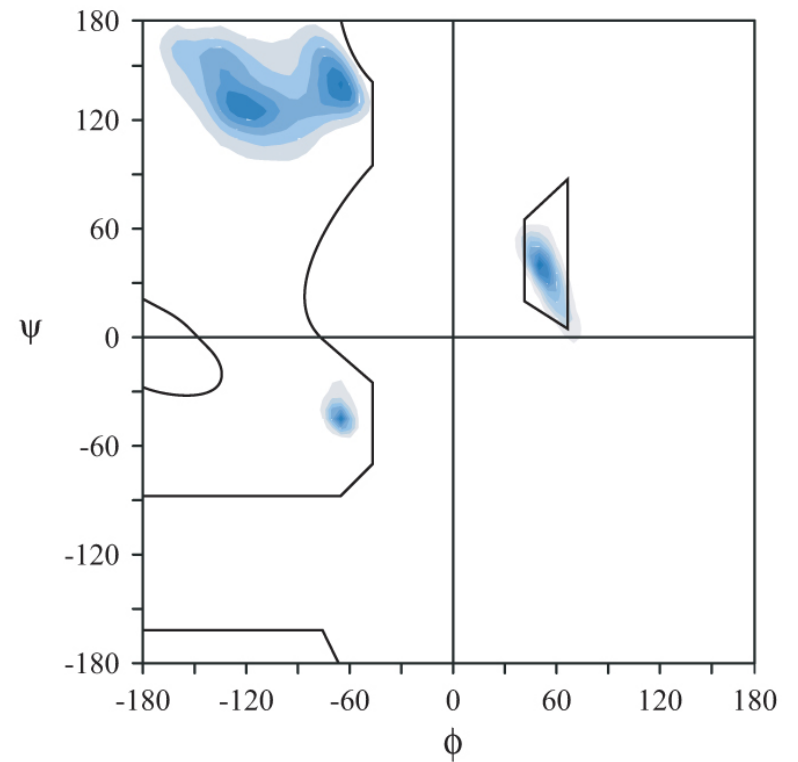
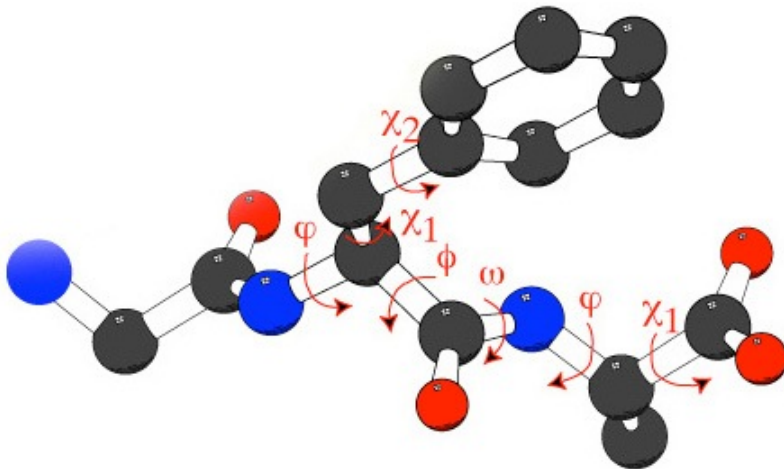
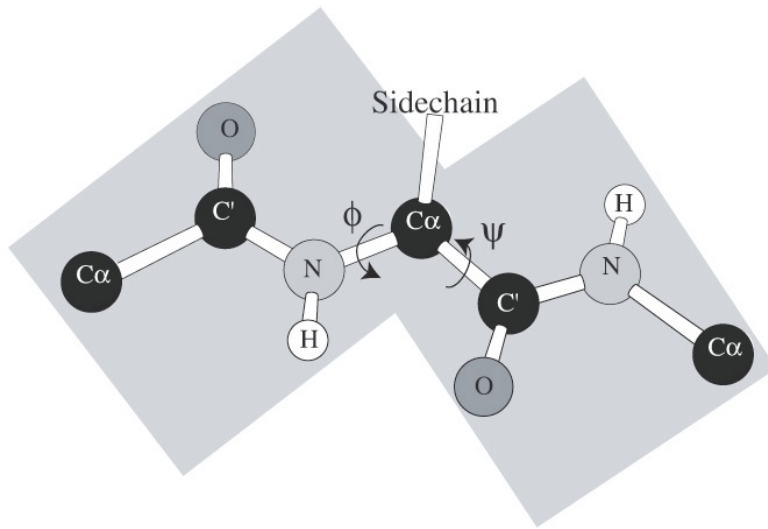
Disadvantages

- At best, only an approximation to reality
- Does not allow precise constructs
- Complex lattices can be as “costly” as the real thing

Non-Lattice Models



Torsional, Not Cartesian



Torsional movements

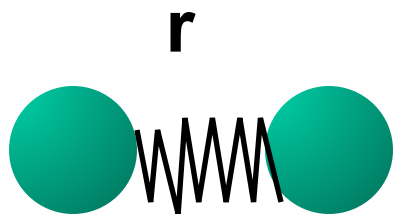
- Advantages:
 - Changes multiple angles simultaneously by using fragments from the library
 - Angular changes are discrete, not continuous
- Disadvantages
 - Small changes disturb the structure
 - Not sampling realistic motions

Energy functions - 2nd problem

Standard Energy Function

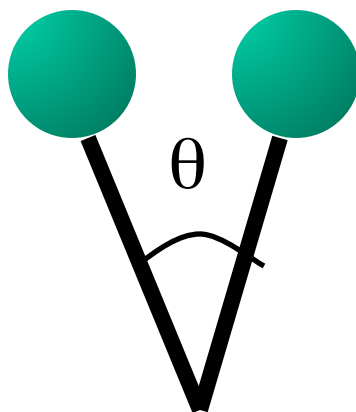
$$\begin{aligned} E = & K_r(r_i - r_j)^2 + & \text{Bond length} \\ & K_\theta(\theta_i - \theta_j)^2 + & \text{Bond bending} \\ & K_\phi(1 - \cos(n\phi_j))^2 + & \text{Bond torsion} \\ & q_i q_j / 4\pi\epsilon r_{ij} + & \text{Coulomb} \\ & A_{ij}/r^6 - B_{ij}/r^{12} + & \text{van der Waals} \\ & C_{ij}/r^{10} - D_{ij}/r^{12} & \text{H-bond} \end{aligned}$$

Energy Terms



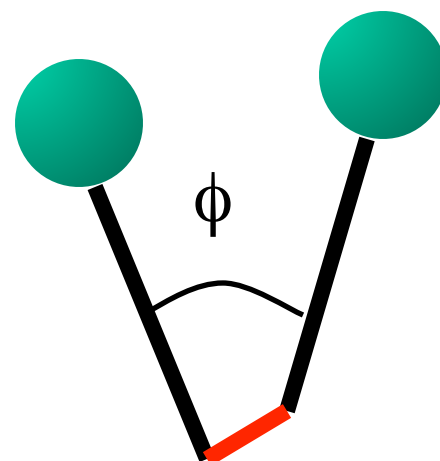
$$K_r(r_i - r_j)^2$$

Stretching



$$K_\theta(\theta_i - \theta_j)^2$$

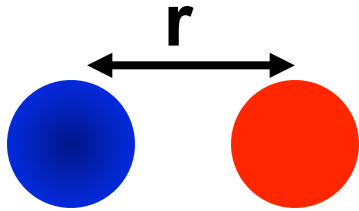
Bending



$$K_\phi(1 - \cos(n\phi_j))^2$$

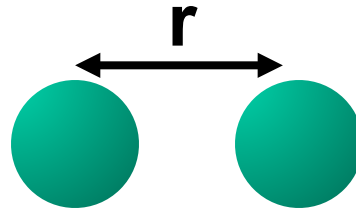
Torsional

Energy Terms



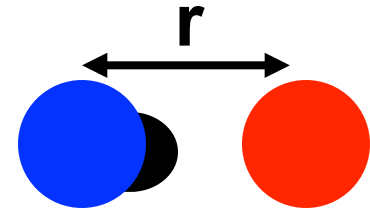
$$q_i q_j / 4\pi\epsilon r_{ij}$$

Coulomb



$$A_{ij}/r^6 - B_{ij}/r^{12}$$

van der Waals



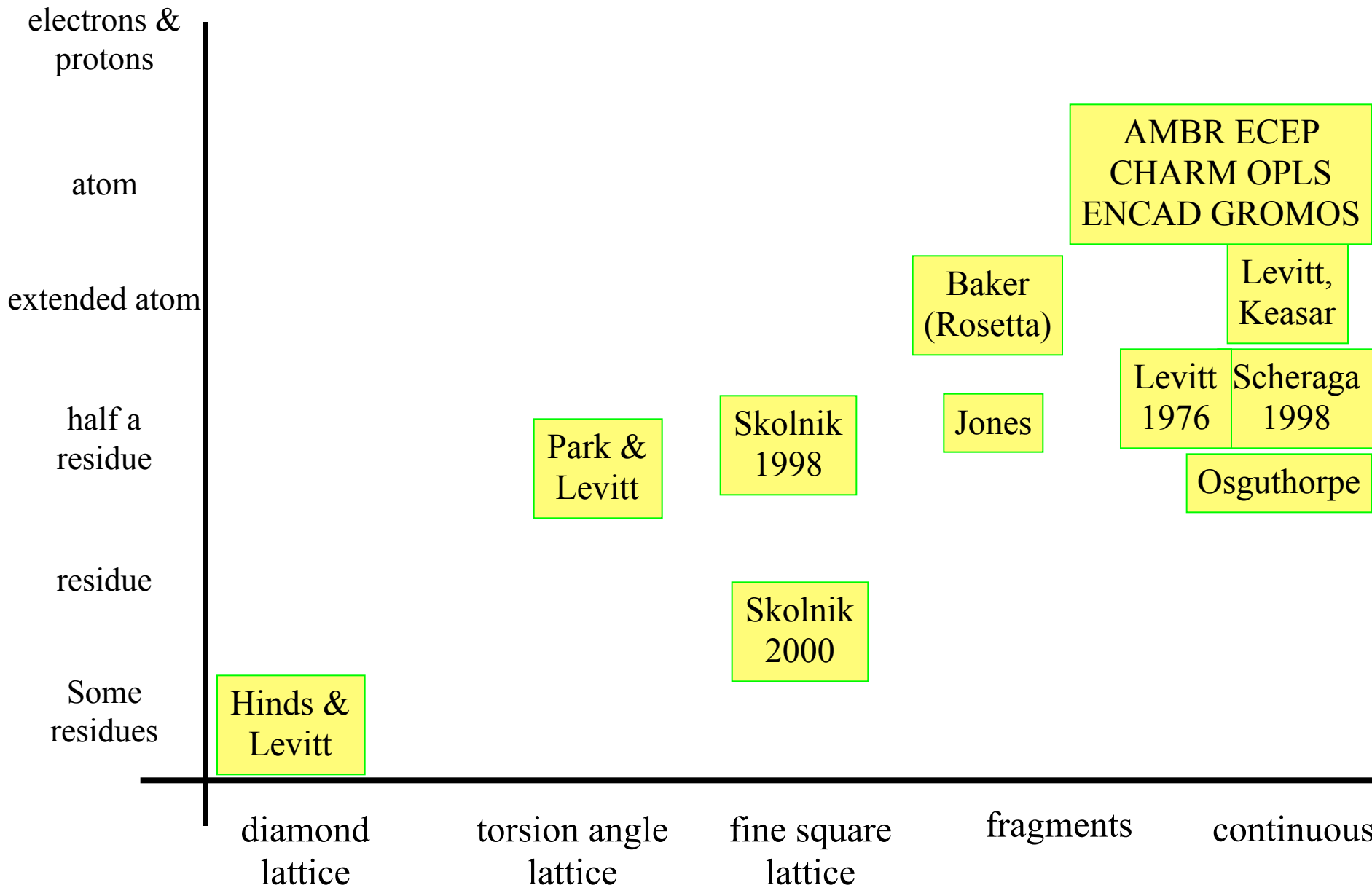
$$C_{ij}/r^{10} - D_{ij}/r^{12}$$

H-bond

Reduced complexity models

- No side chains
 - sometimes no main chain atoms either
 - Or represent the side chain with C_β
- Reduced degrees of freedom
- On-or off-lattice
- Generally have an environment -based score and a knowledge-based residue-residue interaction term
- Sometimes used as first step to prune the enormous conformational space, then resolution is increased for later fine-tuning

Basic element



Protein folding energy landscape

- protein energy landscape is complex, with many local minima
- believed to have a funnel-like shape, with global minimum representing native structure

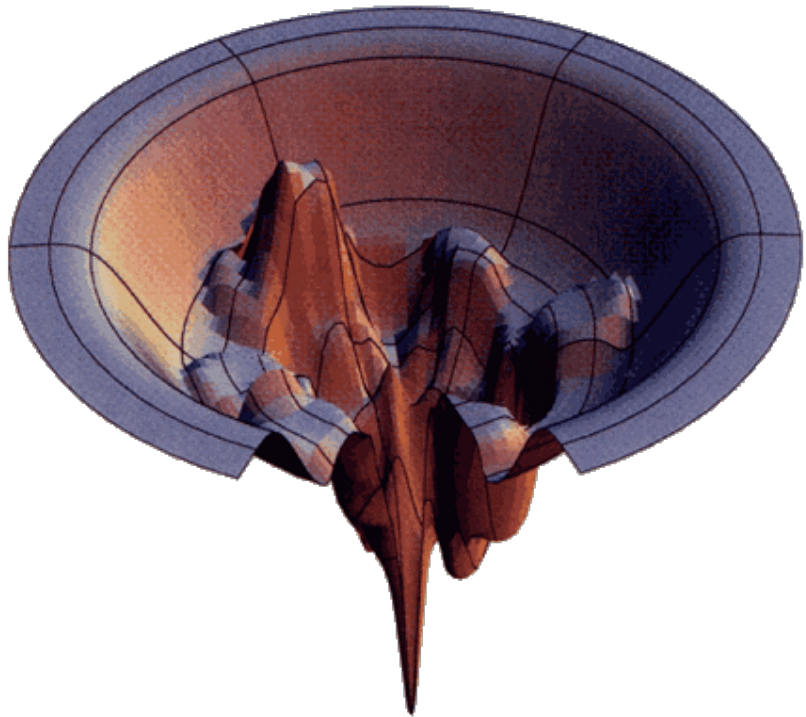


image from <http://bioinfo.mshri.on.ca/>

Problems with energy functions

- Not accurate enough
 - The energy difference between folded/unfolded is often only 5-10 kcal/moles
 - 1000s of energy terms, sum of error is large
- Water
 - For accurate calculation inclusion of water is needed.
 - Implicit water models are quite slow
 - Explicit water needs time to equilibrate

Problems (cont)

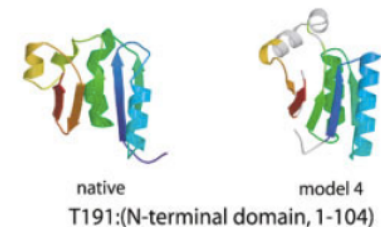
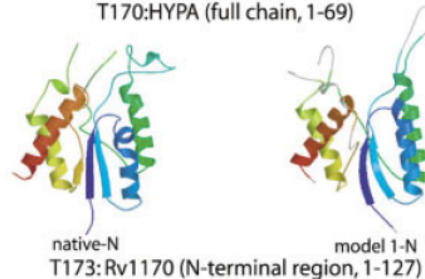
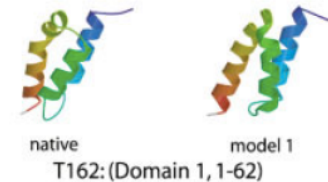
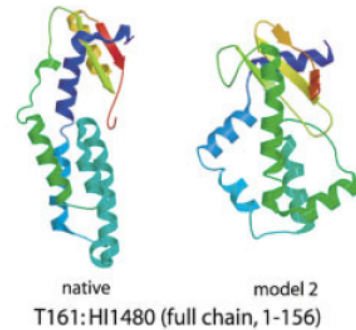
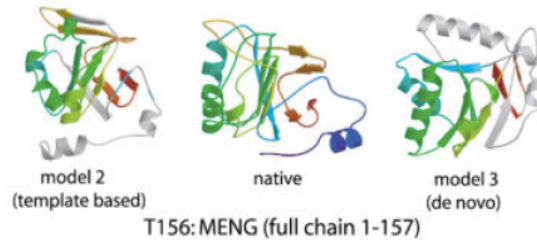
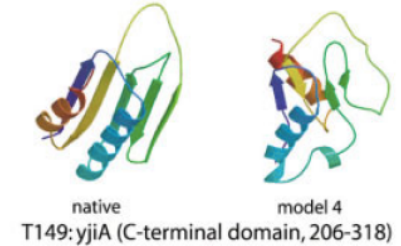
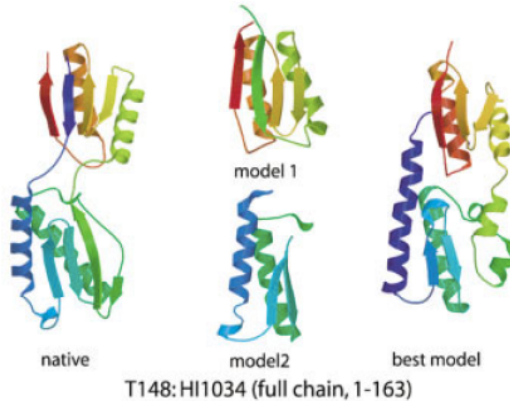
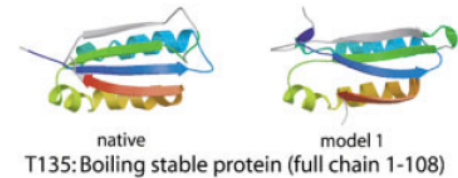
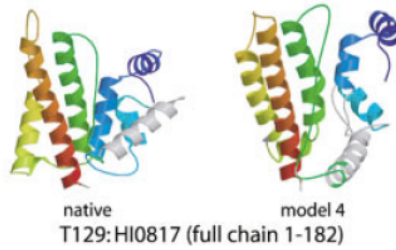
- Entropy
 - We are not searching for the energy minimum, but for the free energy minimum, i.e. MD simulations needed.
- Local minimum problem
 - The barriers are often extremely high to go from one minima to the next.
 - Sidechains cannot pass through each others

Solutions ?

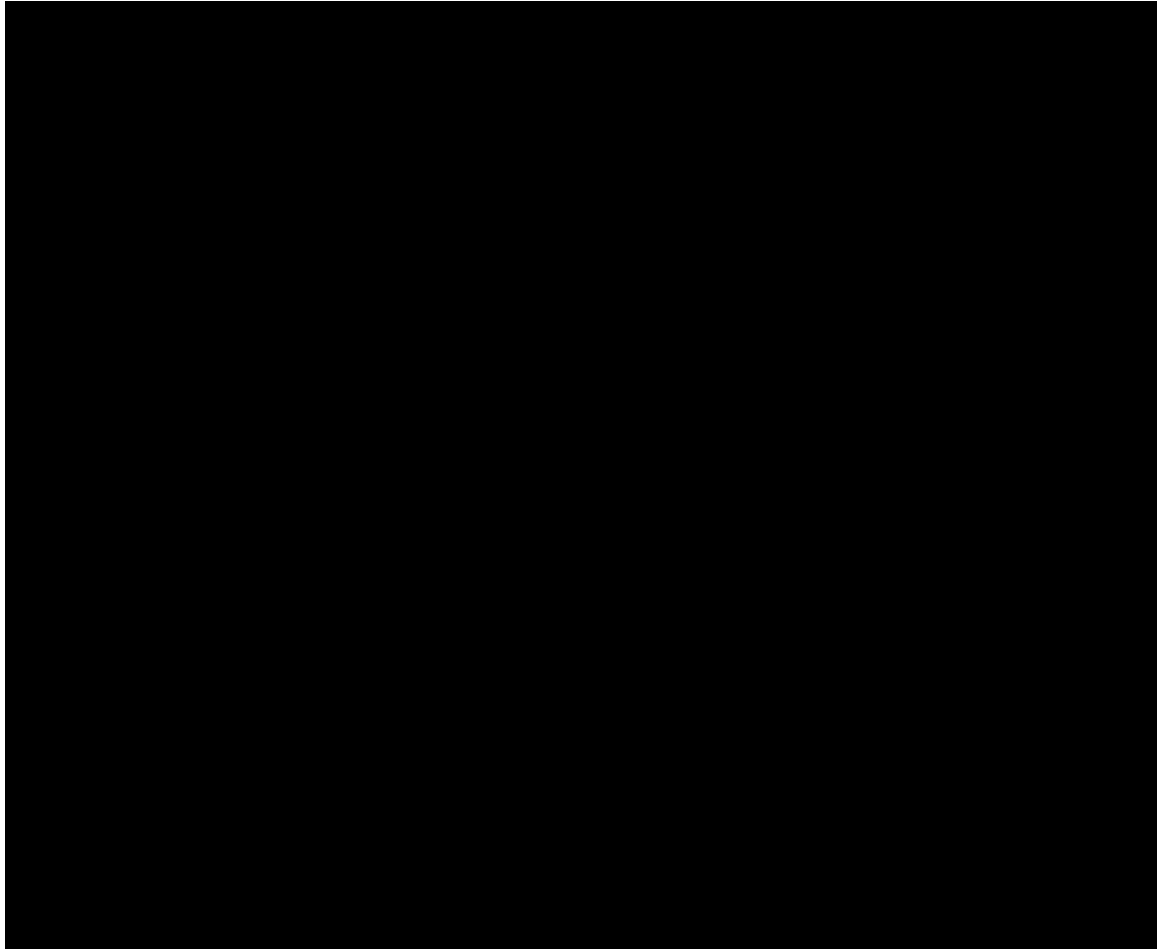
- Are there some ways around these problems
- How does proteins really fold ?
- Can we divide the problem into subproblems ?
 - Local preferences
 - Dominated by sequential information
 - Globular structures
 - Dominated by hydrophobicity
- !!!!! FRAGMENTS !!!!!

Best Method Until
Recently

**Rosetta -
David Baker –
CASP 5
structure
prediction
competition**



Rosetta

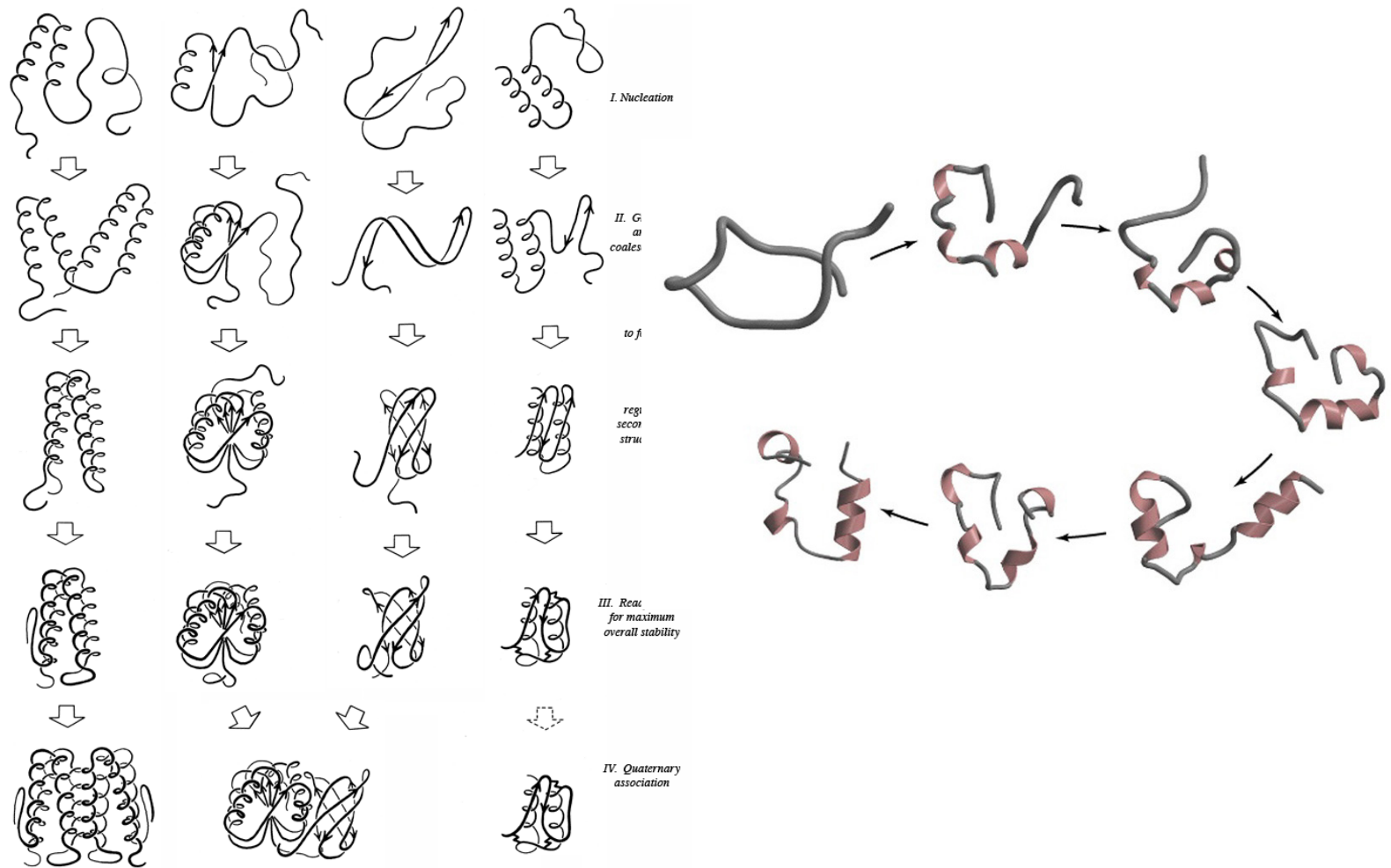


- <http://www.youtube.com/watch?v=GzATbET3g54>

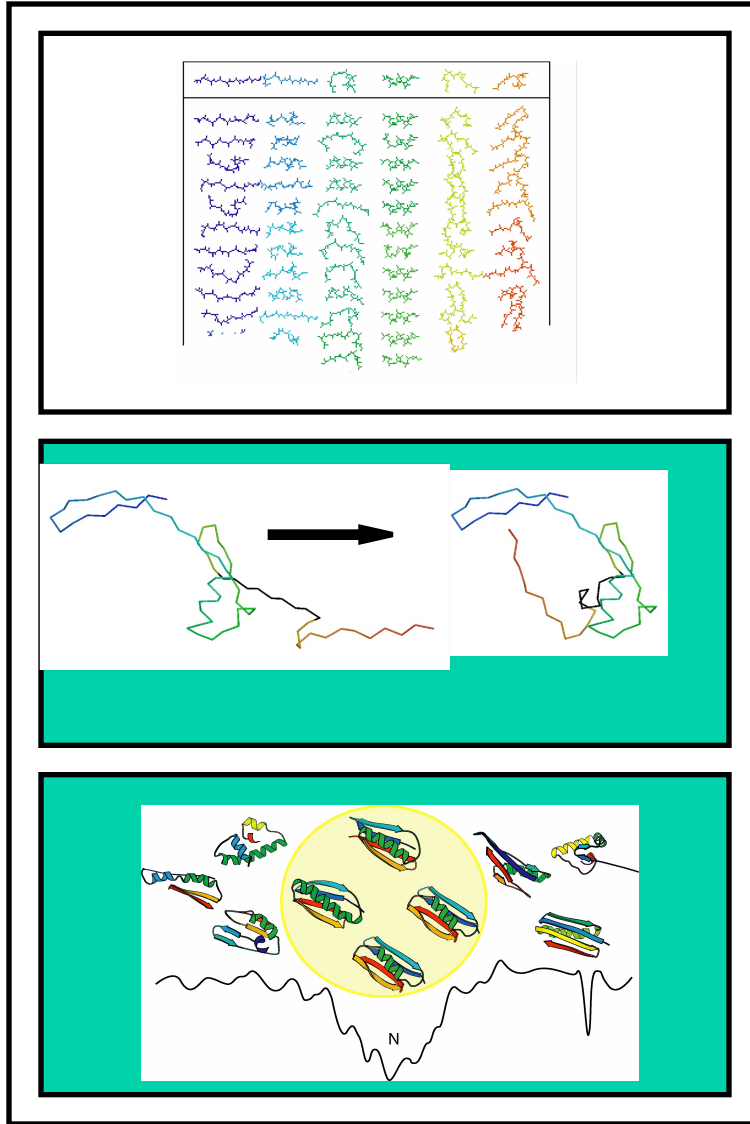
Physics Behind Rosetta

- Proteins are thought to 'collapse' from an unfolded > folded state.
- Local conformations precede and guide global conformations and tertiary structure.
- Local conformations are largely dependent on local sequence, and are finite in number.

Theory Behind Rosetta

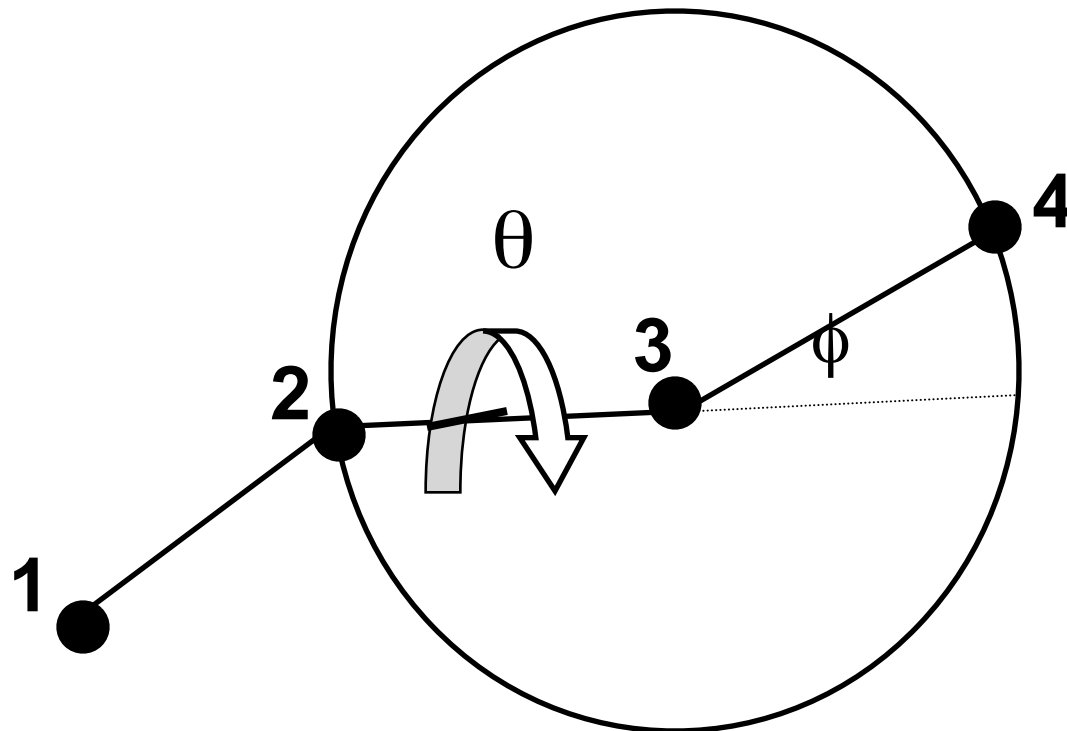


Structure Prediction with Rosetta



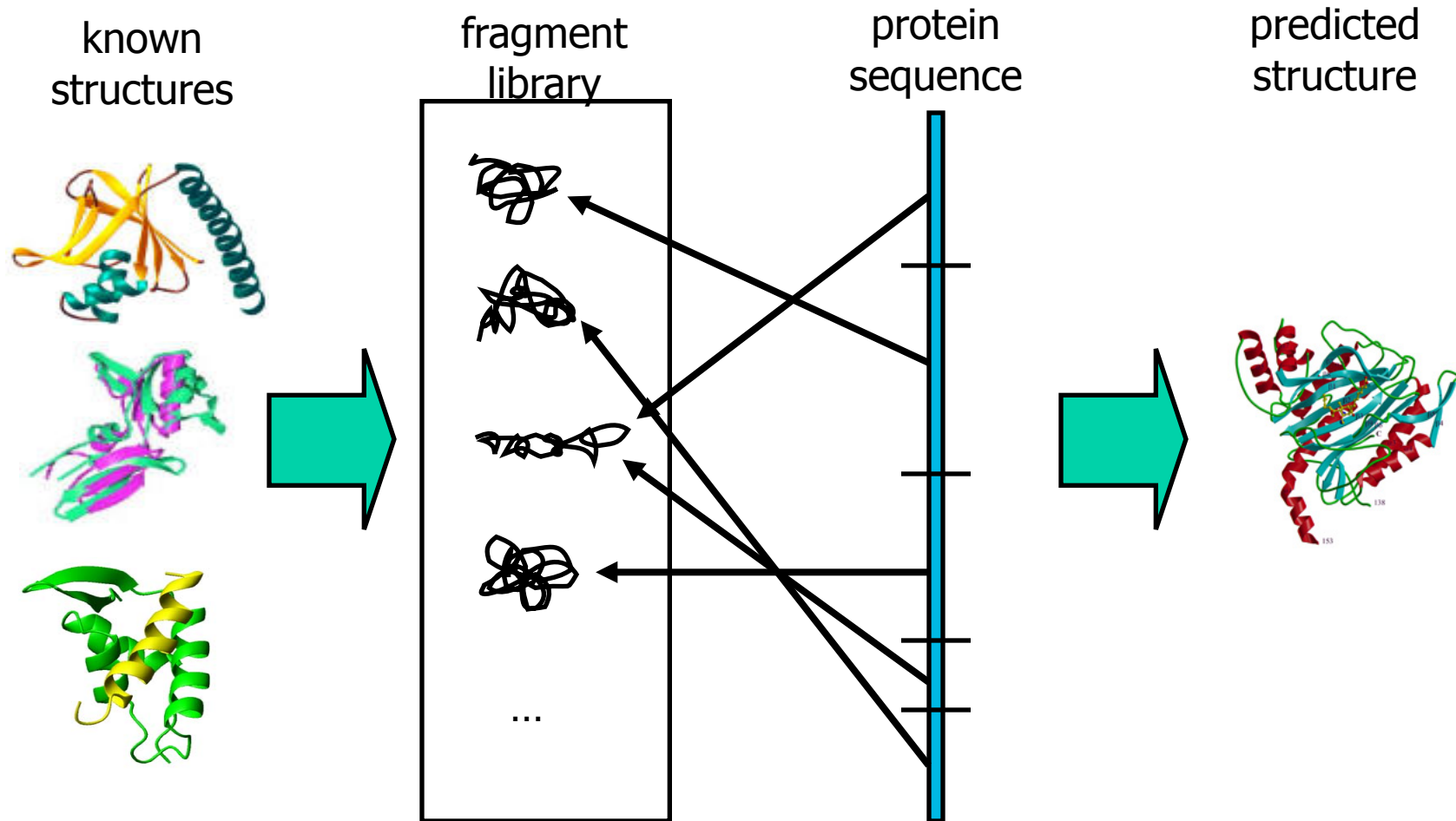
- Select fragments consistent with local sequence preferences
- Assemble fragments into models with native-like global properties
- Identify the best model from the population of decoys

Simplified Chain Representation



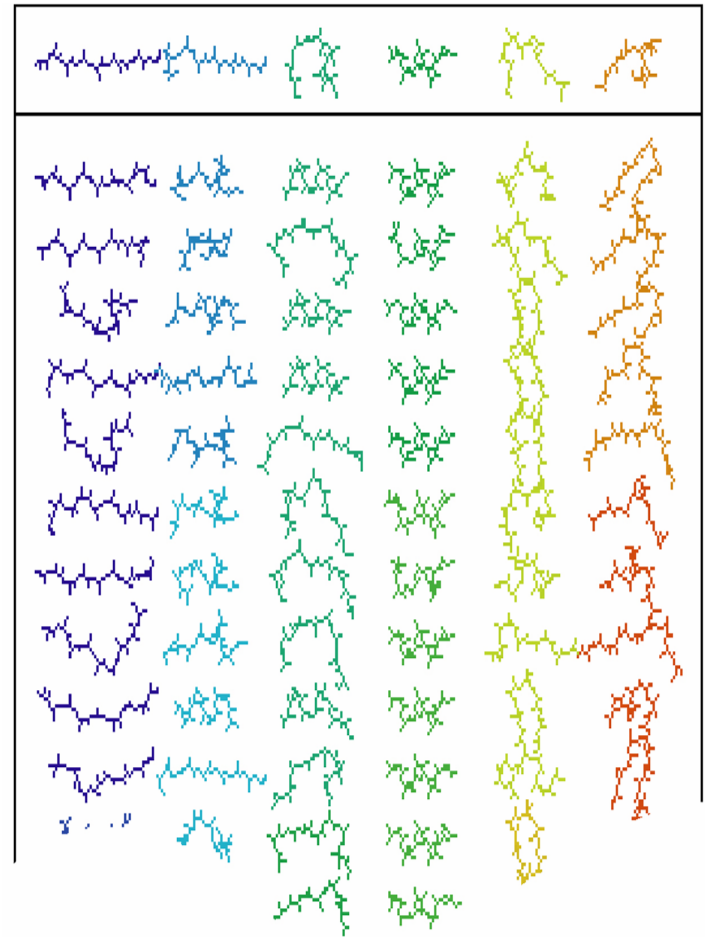
Spherical Coordinates

Assembly of sub-structural units



Build the Fragment Library-Rosetta

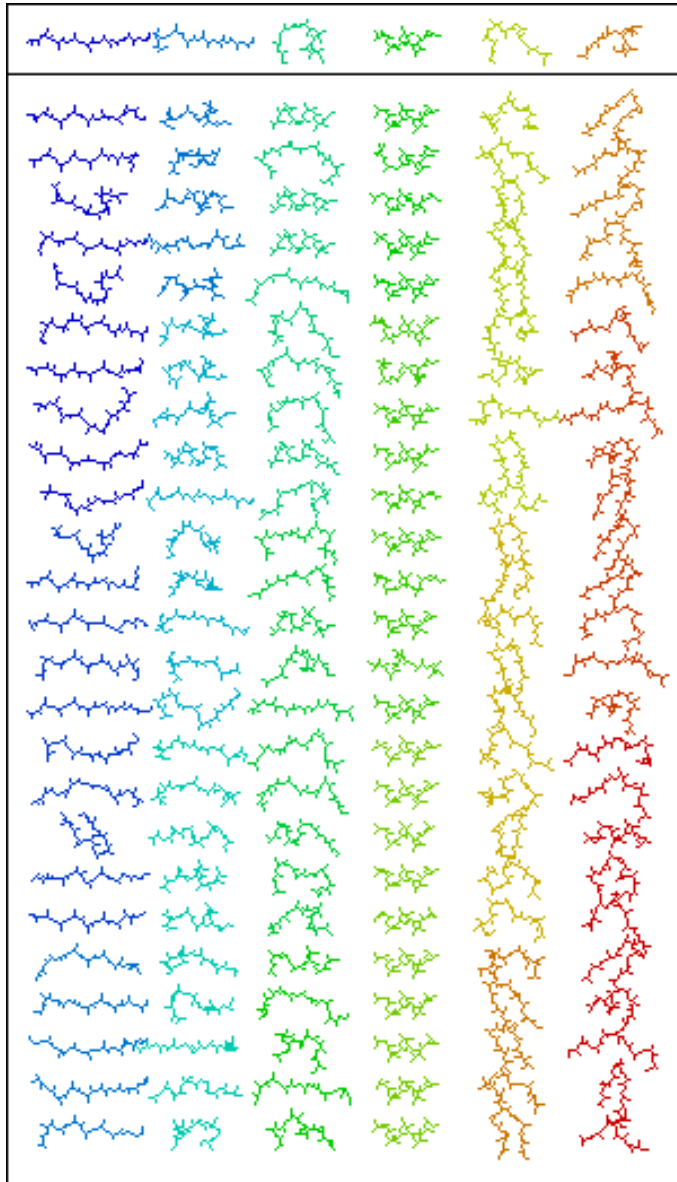
- Extract possible local structures from PDB



Generate the Fragment Library

- Select PDB template
 - Select Sequence Families
 - Each Family has a single known structure (family)
 - Has no more than 25% sequence identity between any two sequence
- Clustering the fragments
 - Generate all the fragments from the selected families

Rosetta Fragment Libraries



- 25-200 fragments for each 3 and 9 residue sequence window
- Selected from database of known structures
 - > 2.5Å resolution
 - < 50% sequence identity
- Ranked by sequence similarity and similarity of predicted and known secondary structure

Scoring Function

- Ideal energy function
 - Has a clear minimum in the native structure.
 - Has a clear path towards the minimum.
 - Global optimization algorithm should find the native structure.

Rosetta MC Energy Function

- Compactness (radius of gyration)
- Hydrophobic burial
- Polar side chain contacts (statistical pairwise potential)
- Hydrogen bonding between beta-strands
- Hard-sphere repulsion (VdW)

Fragment Insertion

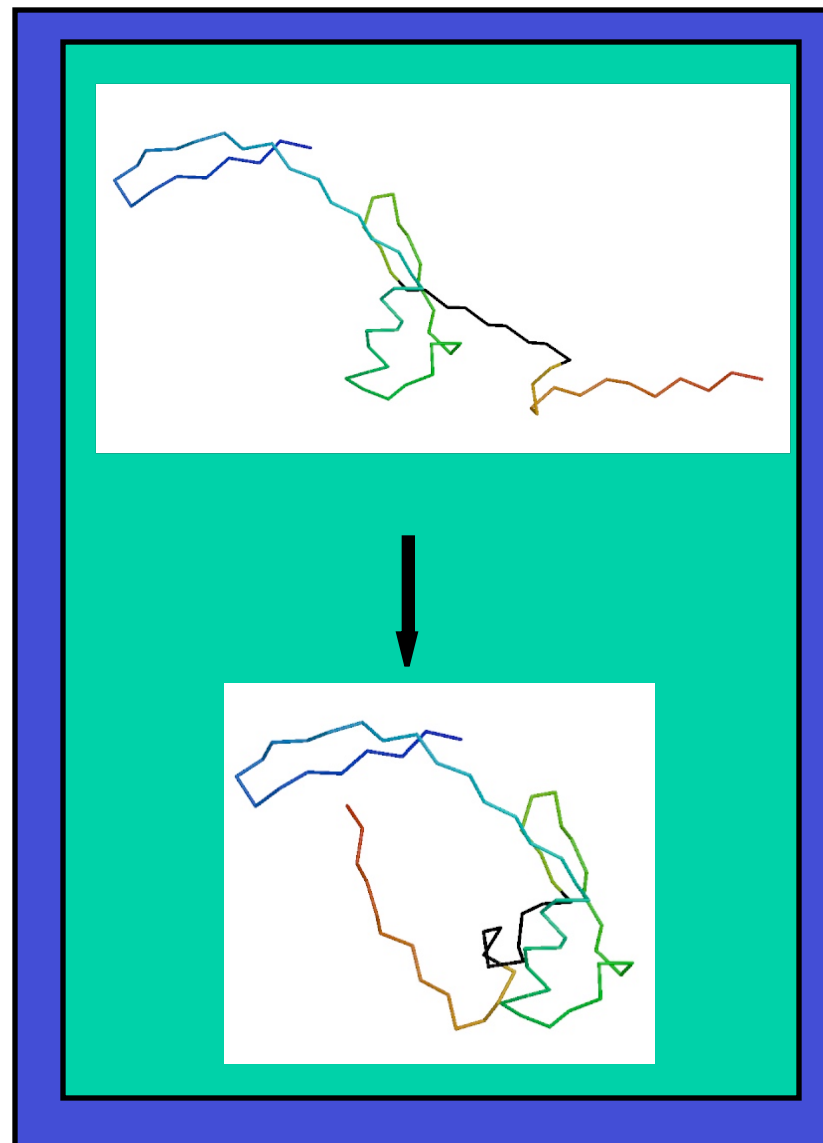
- Finds three and nine residue fragments from known library and replaces unknown torsion angles with the 'known' ones
- Scores all windows of three and nine residues
- Create fragment list with the 200 best three residue and 200 best nine residue fragments

Fragment Assembly

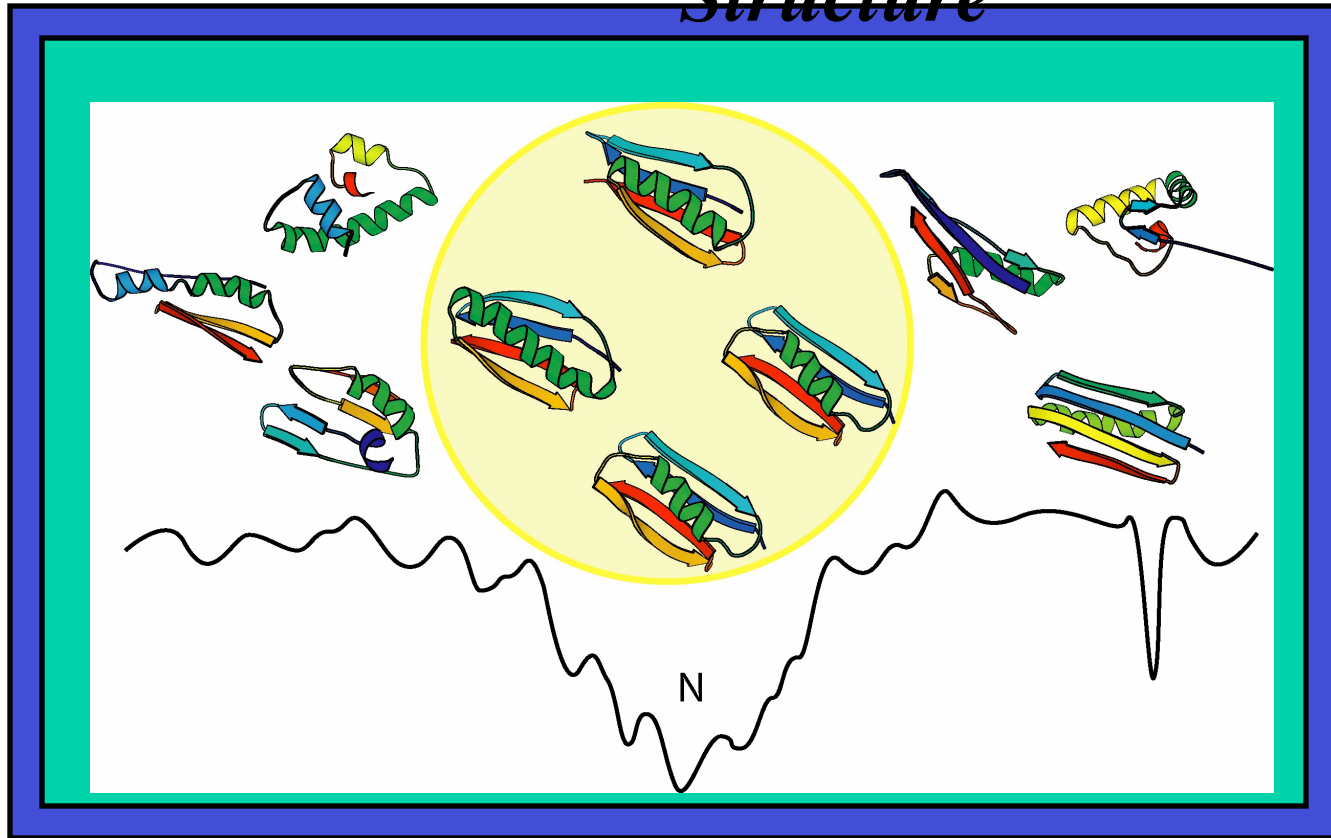
- Randomly choose a nine residue fragment from the top 25 fragments in the ranked list
 - Score this replacement, negatives are kept
- Each simulation chooses a different random start and attempts 28,000 nine residue insertions
- Next 8,000 attempted three residue insertions are scored with the overall structure

Rosetta Potential Function

- Derived from Bayesian treatment of residue distributions in known protein structures
- Reduced representation of protein used; one centroid per sidechain
- Potential Terms:
 - environment (solvation)
 - pairwise interactions (electrostatics)
 - strand pairing
 - radius of gyration
 - C β density
 - steric overlap



Decoy Discrimination: Identifying the Best Structure



- 1000-100,000 short simulations to generate a population of 'decoys'
- Filter population to correct systematic biases
- Fullatom potential functions to select the deepest energy minimum
- Cluster analysis to select the broadest minimum
- Structure-structure matches to database of known structures

Rosetta: clustering the models

- Compare models to each other with RMSD
- Models can come from different family members
- Cutoff varied to give 80-100 members in largest cluster
- The largest clusters are assumed to contain the best structures (attractors in folding space...?)

Rosetta: Filtering the models

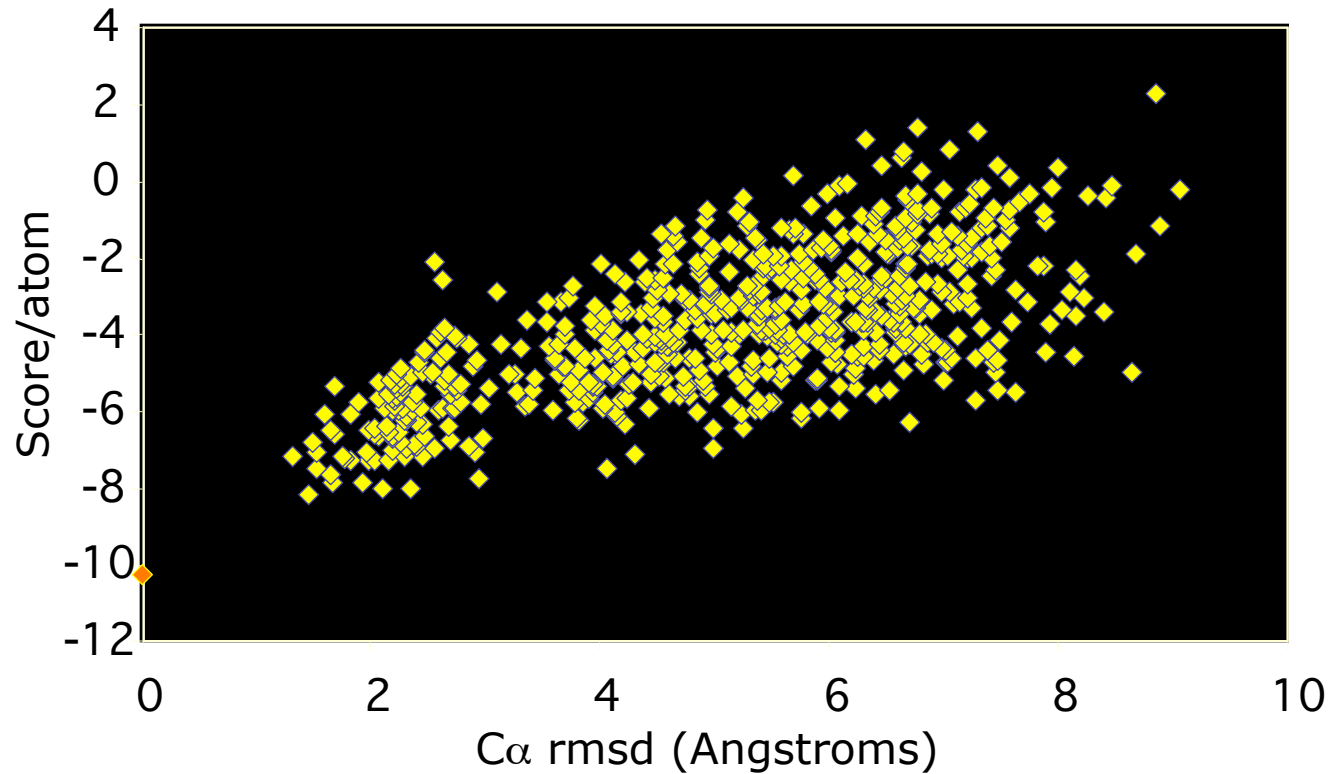
- Between 6,000 and 150,000 models generated
- Contact Order
- Generated models are biased towards simple structures
- Filter models to give correct contact order distribution for domains of that size/composition
- Sheet filter
- Add side chains, calculate atomic physical potential (to eliminate poorly packed structures)

Monte Carlo optimisation

1. Initial configuration (random or extended)
 2. Make a randomised MOVE on configuration
 3. Measure change in quality of structure (DE)
 4. IF better () ACCEPT MOVE
 5. ELSIF rand ACCEPT MOVE
 6. ELSE REJECT MOVE
- GO TO 2. (reduce T if you like)

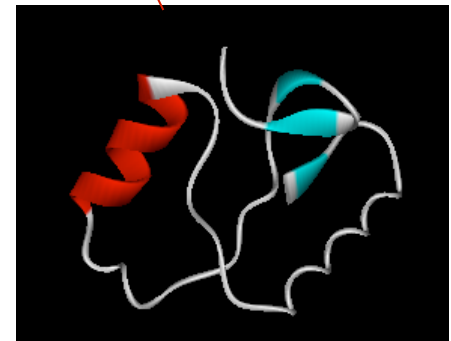
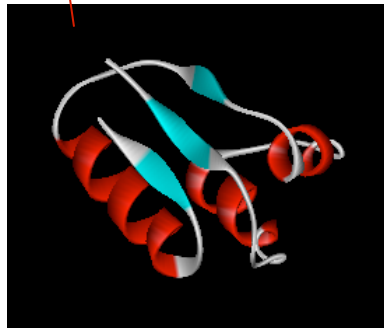
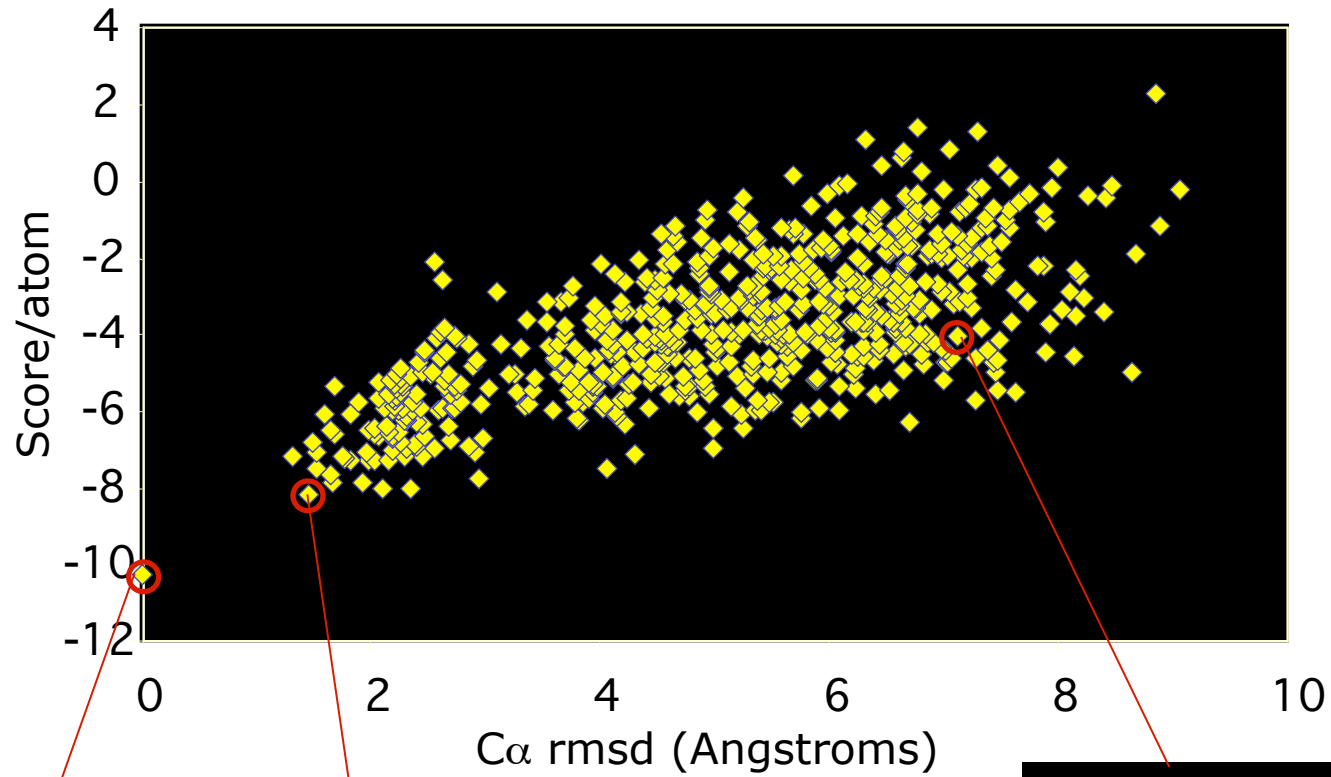
Testing of scoring functions

Contact scores for 1ctf decoy set (4state decoys)

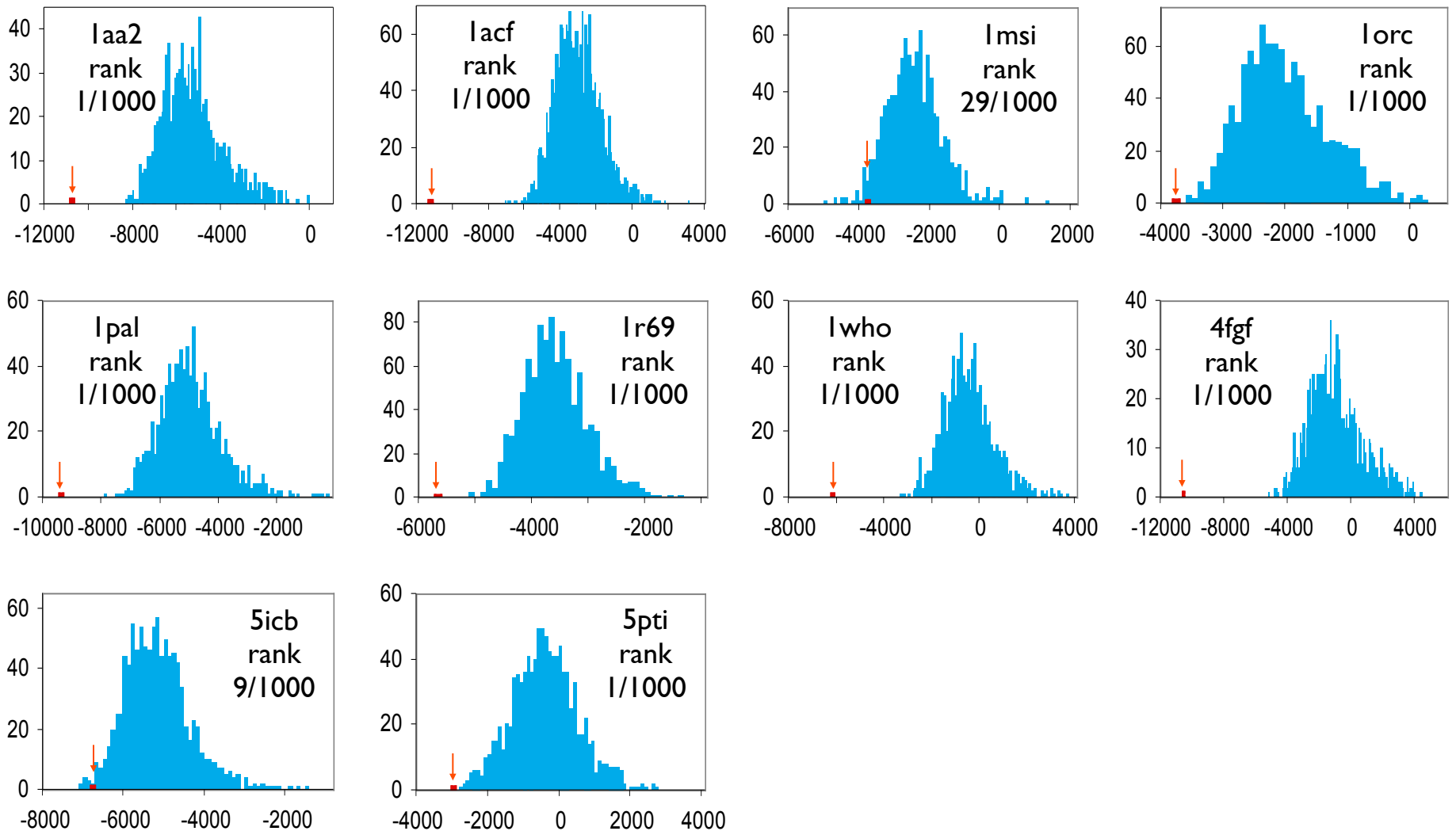


Testing of scoring functions

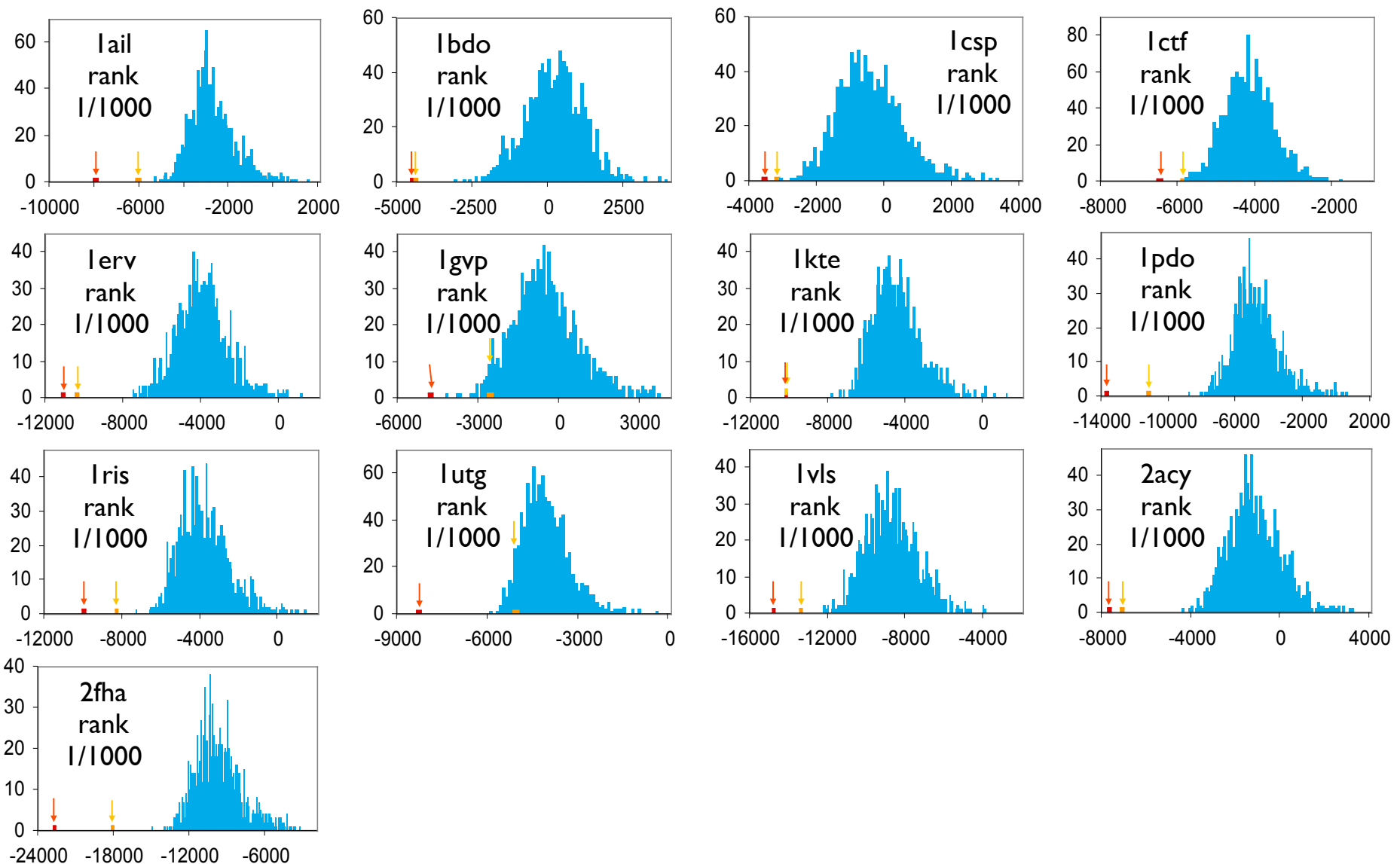
Contact scores for 1ctf decoy set (4state decoys)



Histograms of native (red) and decoy (blue) scores for the Rosetta decoy monomers

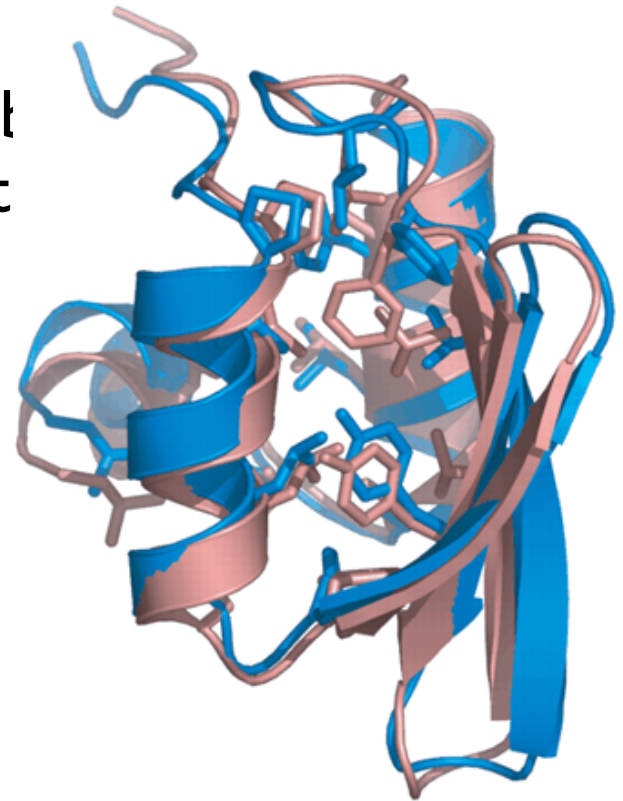


Histograms of native (red) and decoy (blue) scores for the Rosetta decoy oligomers



HR protocol to Rosetta

- Additional refinement step from I clusters using all atom refinement
 1. Make small dihedral changes
 2. Rebuild sidechains
 3. Minimize (in dihedral space)
 4. Evaluate energy
 5. Go To 1
- 5 out of 6 small proteins < 1.5 Å



Fold.It - The protein Folding Game

Solve Puzzles for Science | Foldit

http://fold.it/portal/

Solve Puzzles for Science | Foldit

01:09:06 GMT

foldit BETA
Solve Puzzles for Science

BLOG PUZZLES FEEDBACK GROUPS FORUM PLAYERS WIKI FAQ RECIPES ABOUT CONTESTS CREDITS

Click to learn how you contribute to science by playing Foldit.

GET STARTED: DOWNLOAD

Win Beta
Windows XP/Vista/7

Mac Beta
Intel OSX 10.4 or later

Linux Beta
Linux

RECOMMEND FOLDIT

Send

USER LOGIN

Username: *

Password: *

Log in

- Create new account
- Request new password
- Sign in using Facebook

Connect with Facebook

What's New

Vote for Foldit!

Foldit is in the running for winning NSF's International Science & Engineering Visualization Challenge! You can vote for Foldit for the People's Choice award in the here: <https://nsf-scivis.skild.com/skild2/NationalScienceFoundation/viewEntryD...> (click "vote for this entry").

(Mon, 10/24/2011 - 22:55 | 0 comments) Share

Small update

We've just posted an update which will let us run Electron Density puzzles again. Also included are some minor bug fixes.

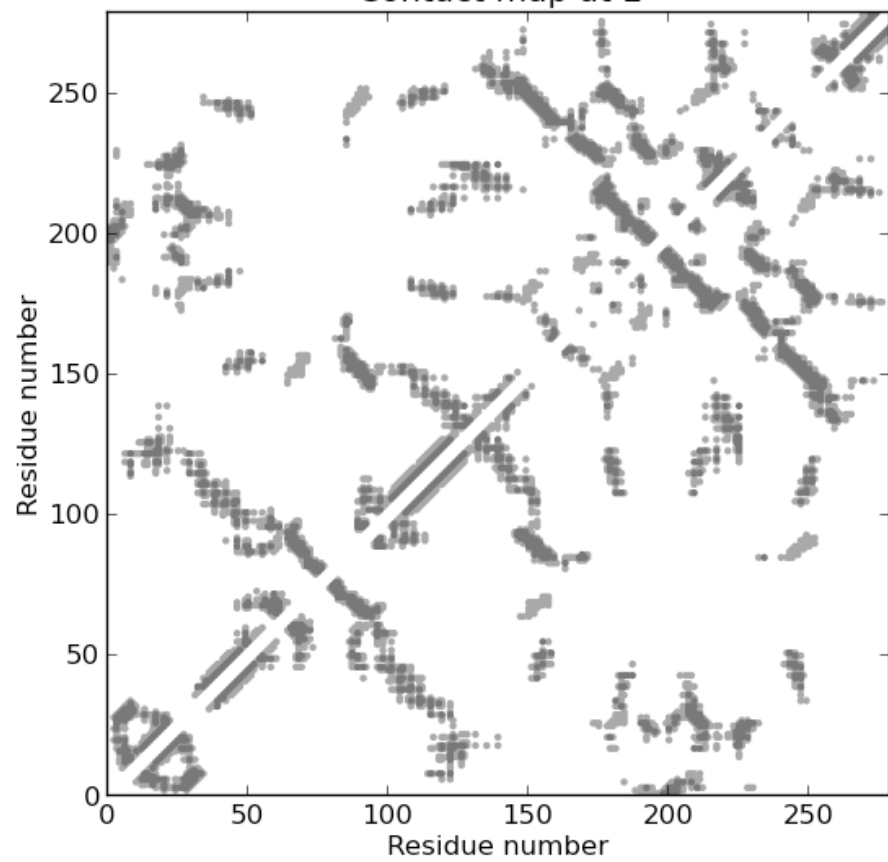
(Wed, 10/19/2011 - 19:15 | 8 comments) Share

Latest Foldit paper named "Article of the month" by Nature Structural & Molecular Biology

SOLOISTS **EVOLVERS** **GROUPS** **TOPICS**

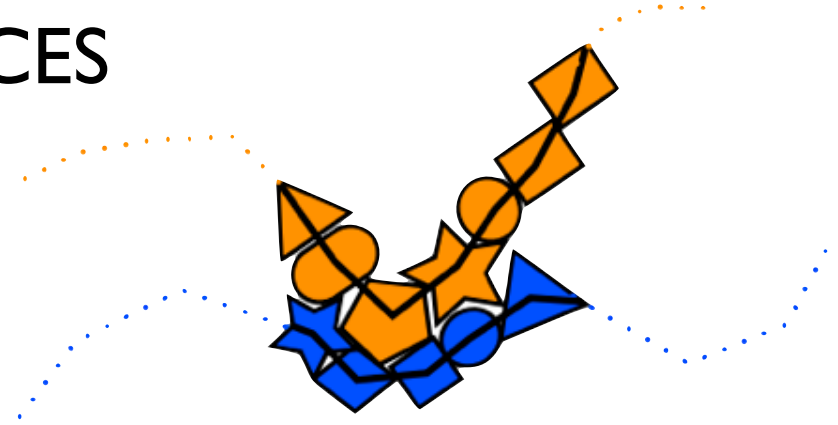
PLAYER	PUZZLE	SCORE
Dr. Goochie 160 512	472: Electron D...e 4	10,839
anthunk 41 12	471: Loop Remod...e 1	8,246
spmm 24 9	470: Revisiting...153	10,858
DarkTigrou 160 19016	470 (<15): Revi...153	10,489
tigatzl 160 846	470 (<150): Rev...153	10,551
wudoo 77 1	469: De-novo Fr... 14	10,442
... 160 178	Beginner Puzzle..._als	9,627

Contact map at L

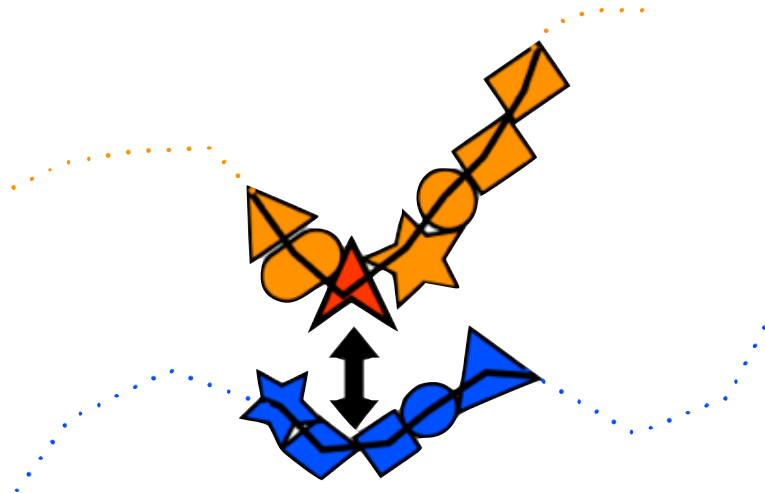


SPATIAL PROXIMITY INDUCES SEQUENCE COUPLING

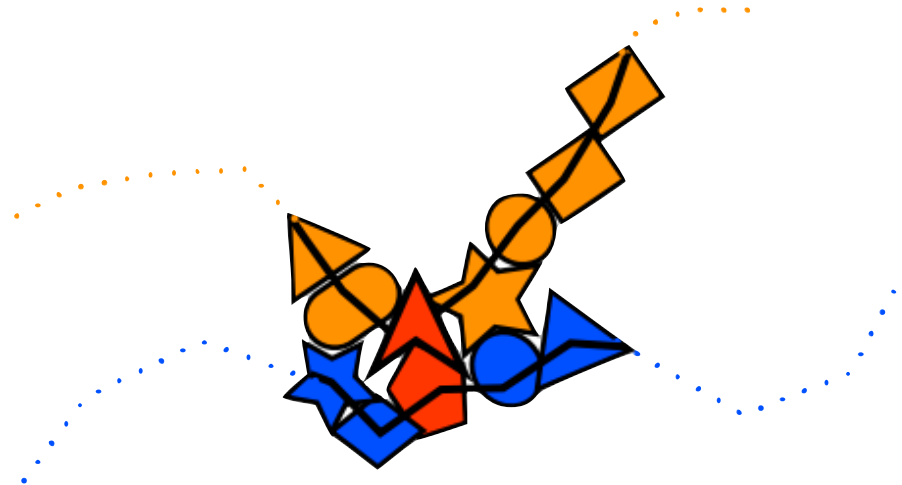
Native interactions

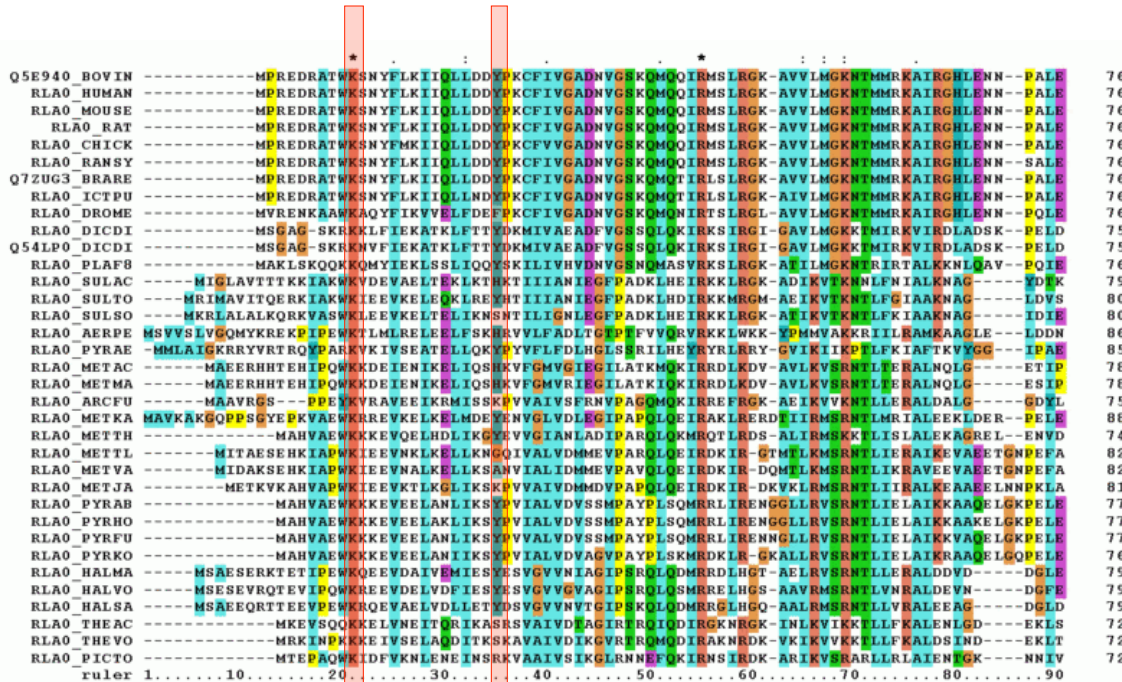


Unfavorable mutation

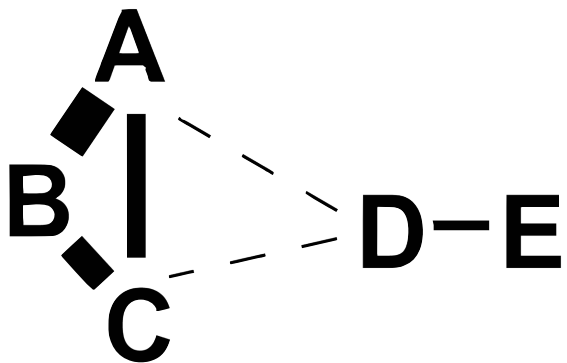


Compensating mutation





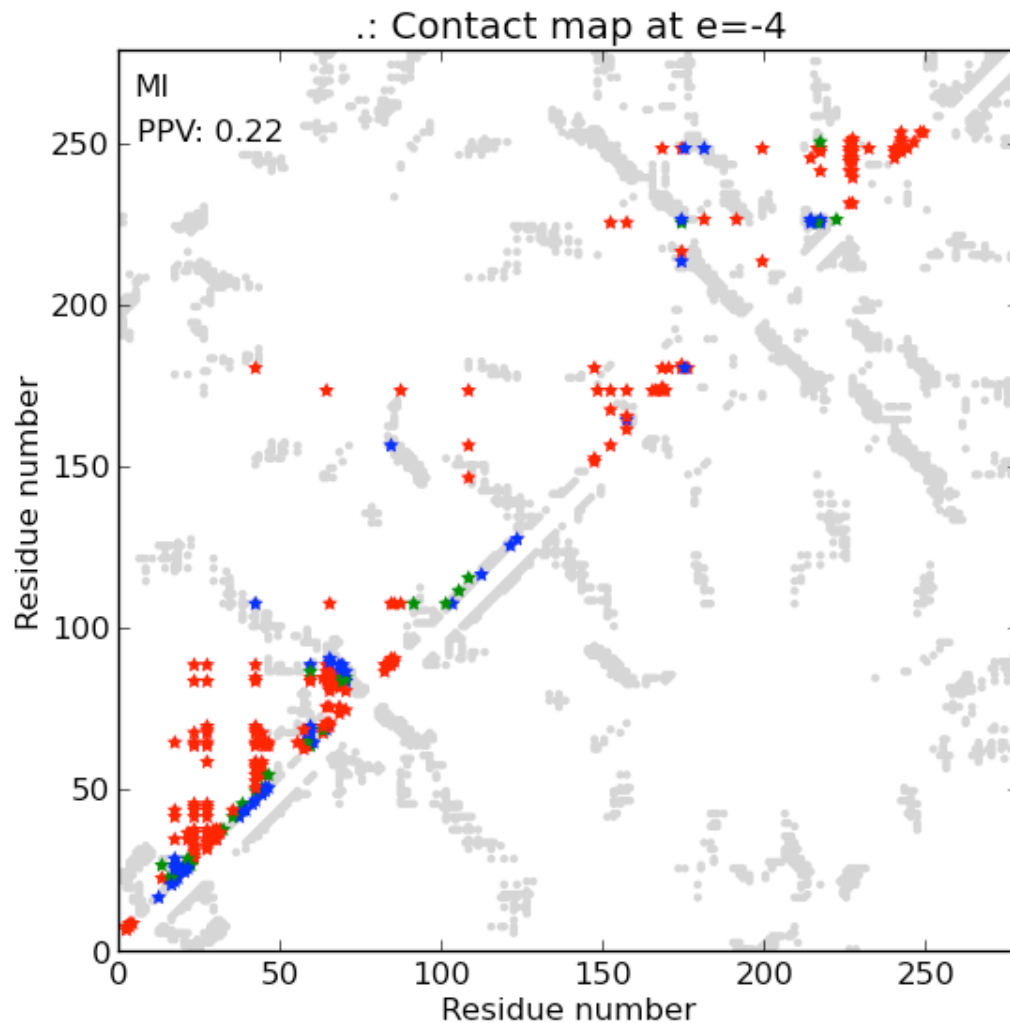
Residues in contact
tend to co-evolve



Decoupling direct interactions
from indirect ones

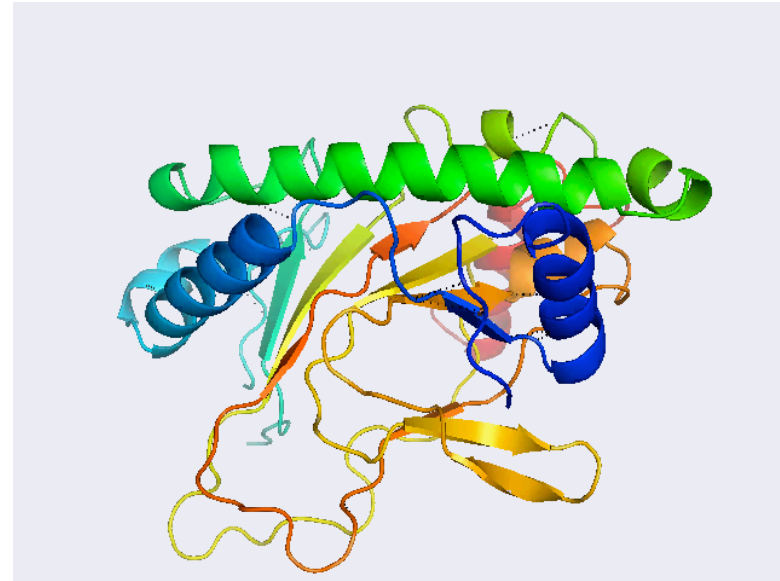
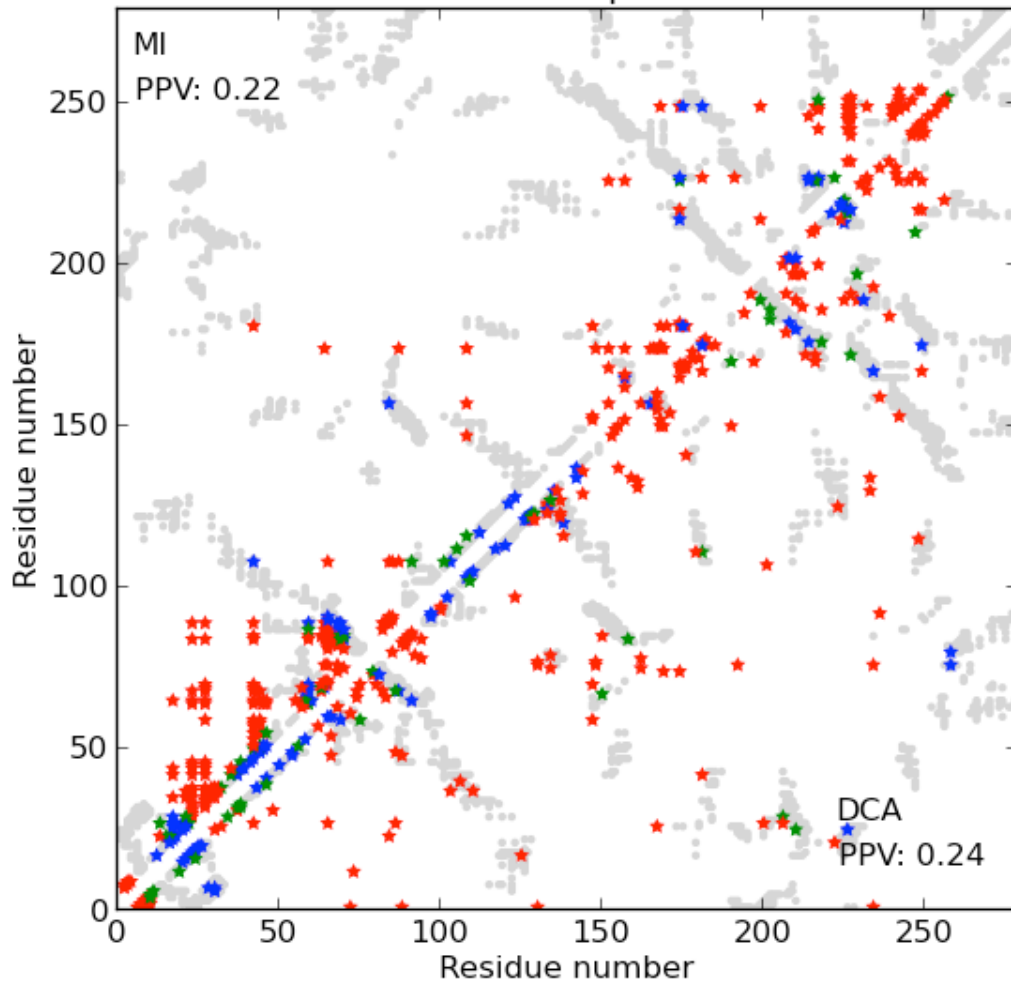
Burger & van Nimwegen (2010) doi:10.1371/journal.pcbi.1000633

MUTUAL INFORMATION

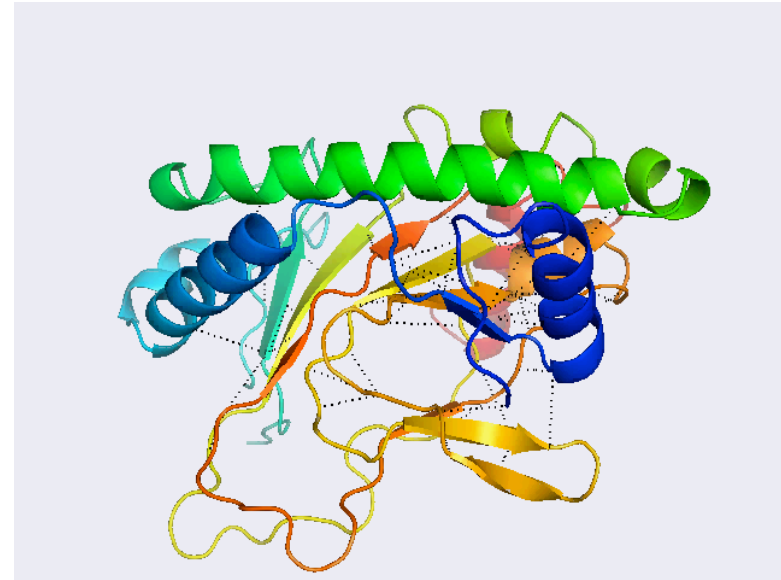
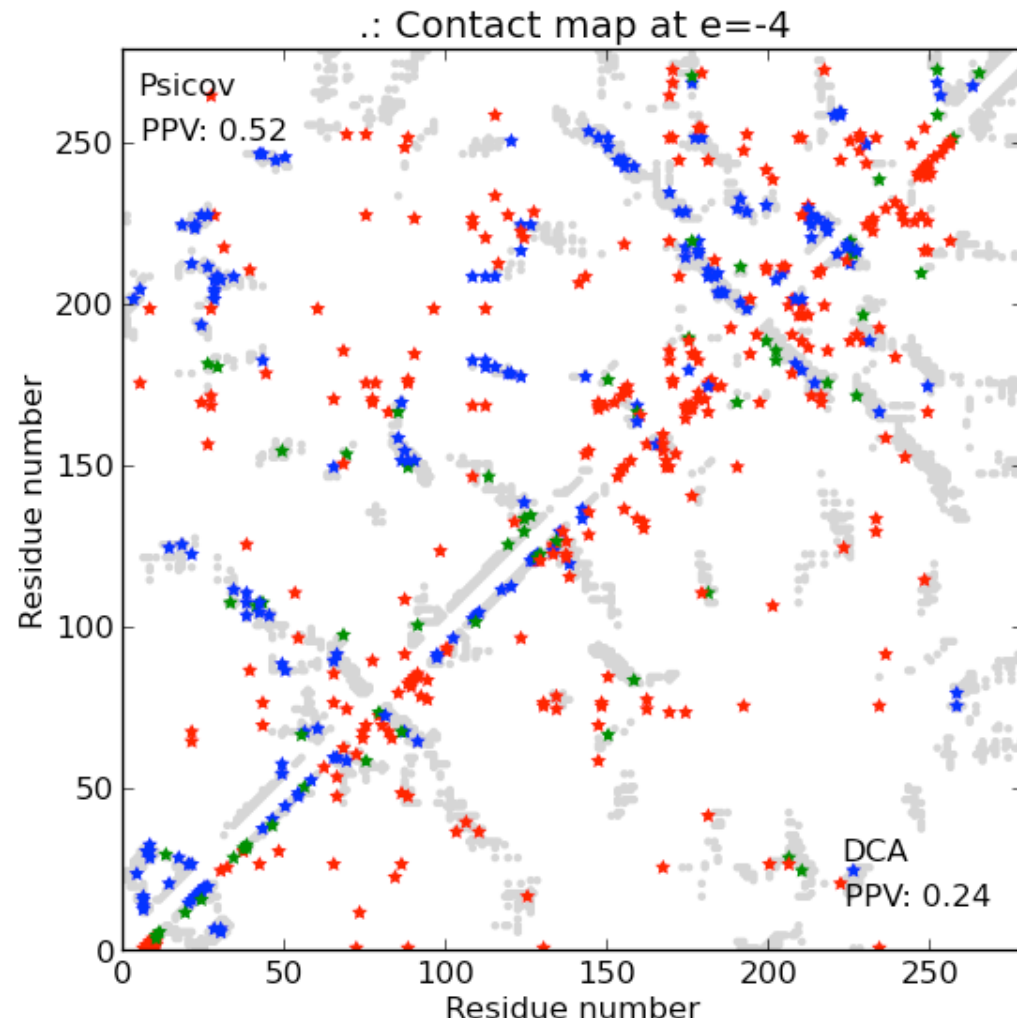


MFDCA

∴ Contact map at $e=-4$



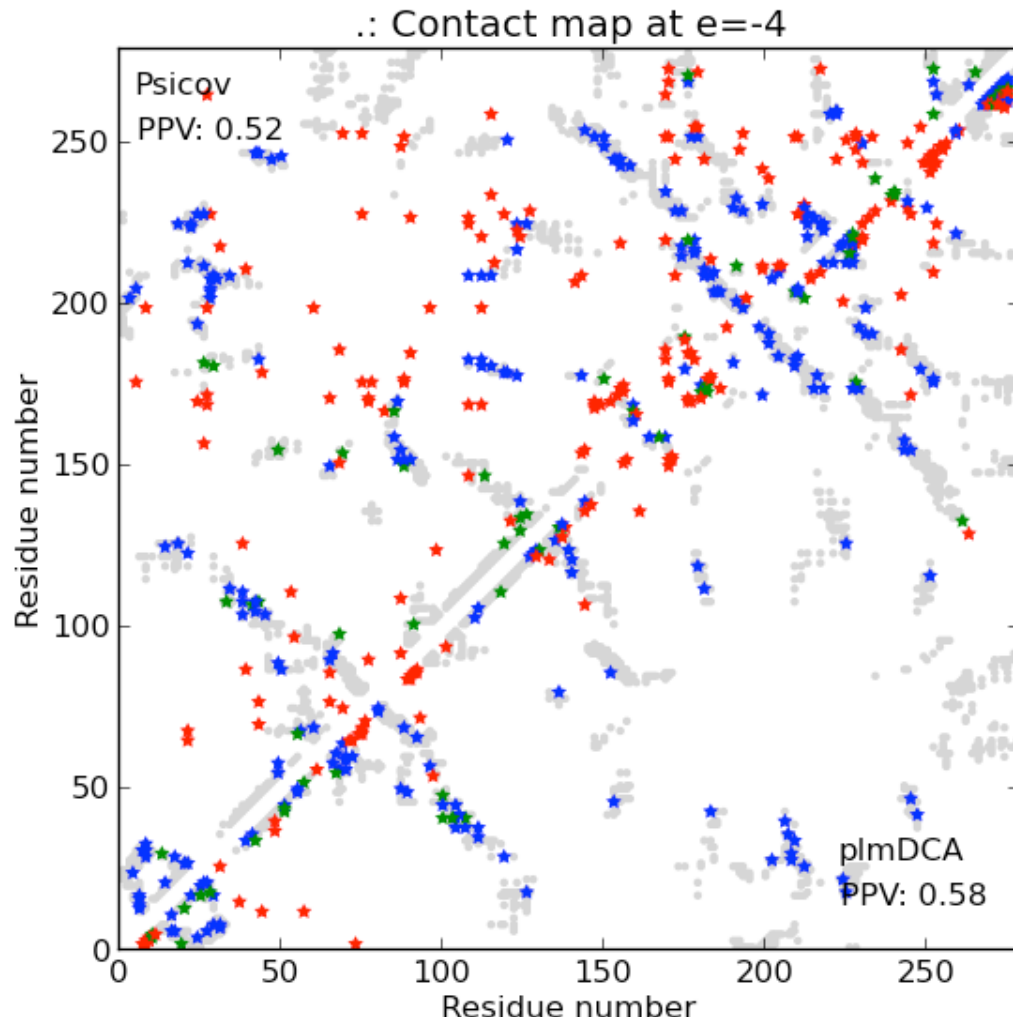
PSICOV

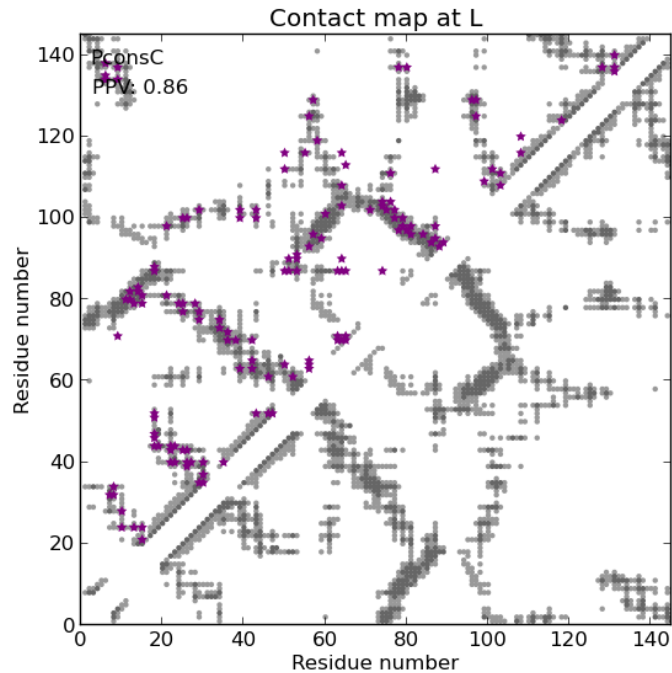


Jones DT, Buchan D.W.A, Cozetto D., Ponti M.

PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments (2012)

PLMDCA





Mannitol-specific phosphotransferase enzyme IIA component

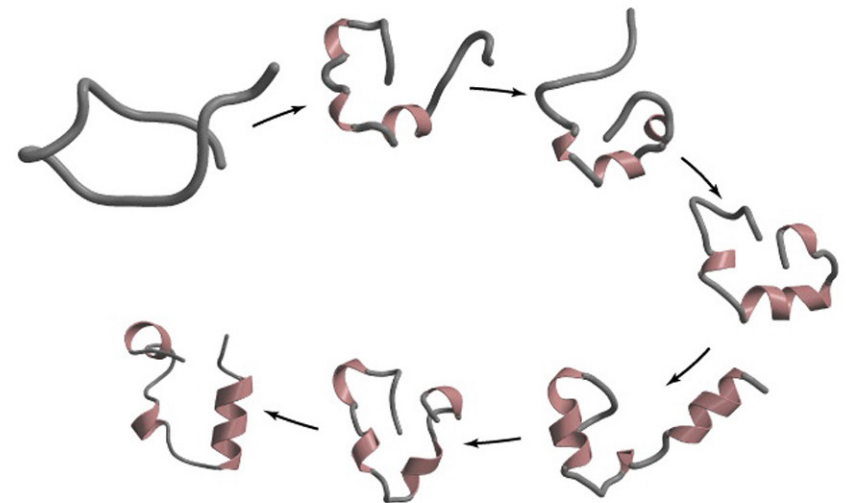
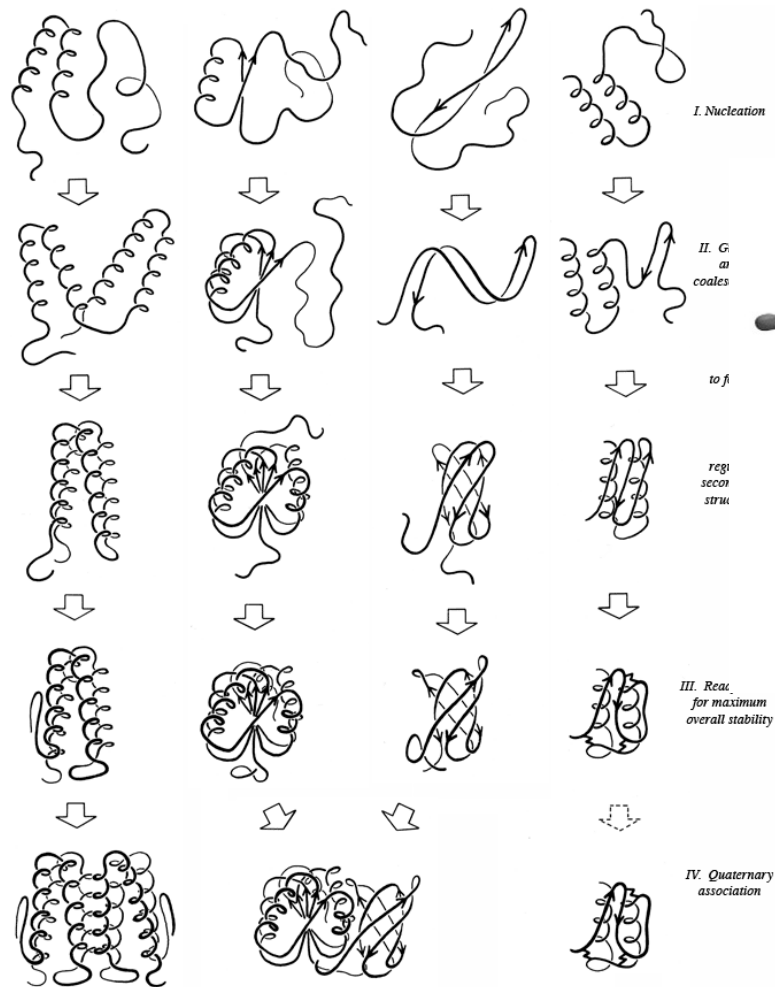
GDT-TS = 0.81

RMSE=2.08 Å

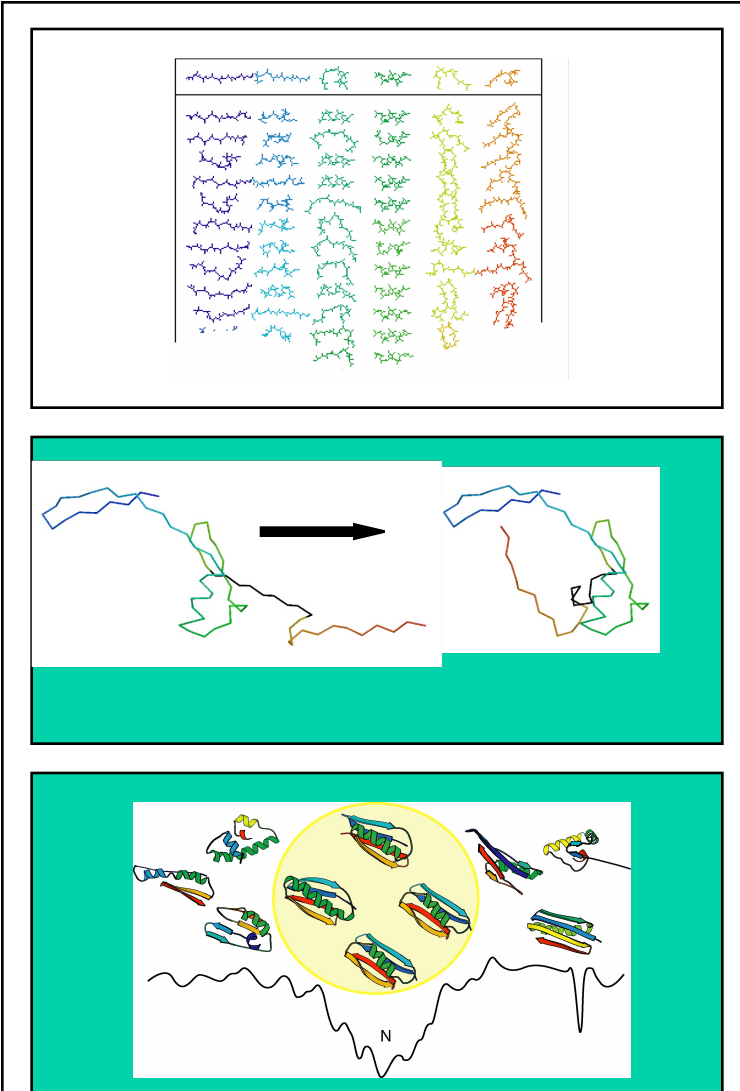
Theory Behind Rosetta

- Proteins are thought to 'collapse' from an unfolded > folded state.
- Local conformations precede and guide global conformations and tertiary structure.
- Local conformations are largely dependent on local sequence, and are finite in number.

Theory Behind Rosetta

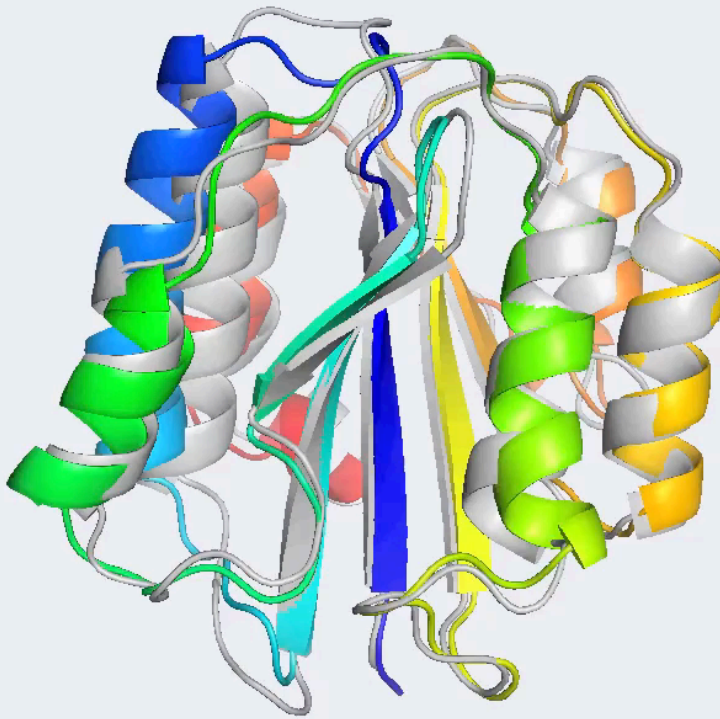


Structure Prediction with Rosetta



- Select fragments consistent with local sequence preferences
- Assemble fragments into models with native-like global properties
- Identify the best model from the population of decoys

EXAMPLE: FOLDING IATZ:A



Similarity to the
native structure:

RMSD	1.75
TM-	0.91
GDT-	0.84

Using PconsC contacts and ROSETTA-based folding protocol