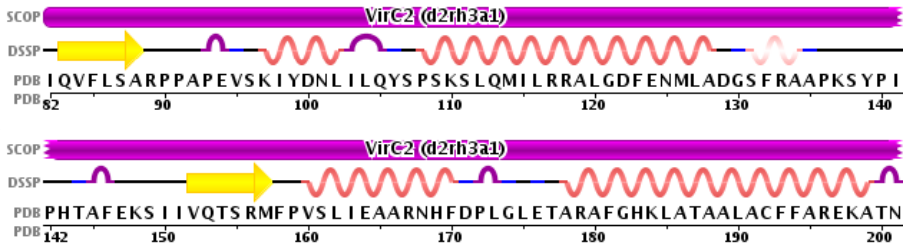
- L12 - Introduction to Protein Structure; Structure Comparison & Classification
- L13 - Predicting protein structure
- L14 - Predicting protein interactions
- L15 - Gene Regulatory Networks
- L16 - Protein Interaction Networks
- L17 - Computable Network Models

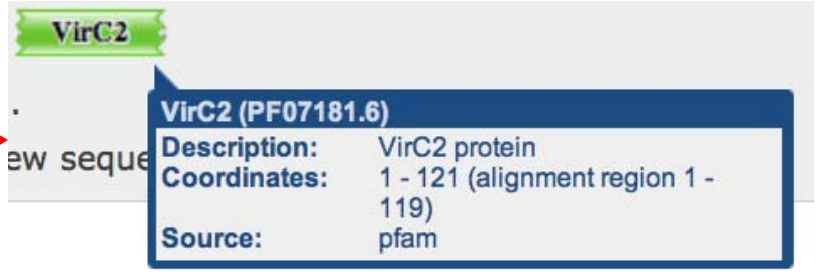
Predicting Protein Structure

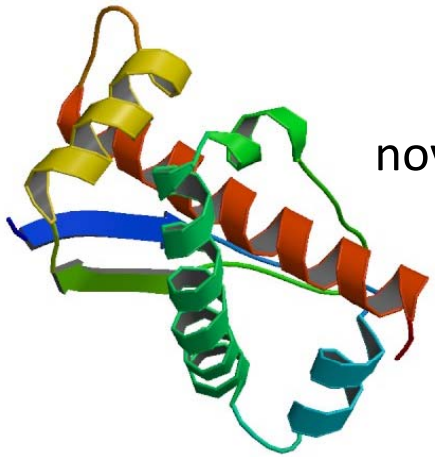
IQVFLSARPPAPEVSKIY
DNLILQYSPSKSLQMILR
RALGDFENMLADGSFR
AAPKSYPIPHTAFEKSIIV
QTSRMFPVSLIEAARNH
FDPLGLETARAFGHKLA
TAALACFFAREKATNS

secondary structure



domain structure



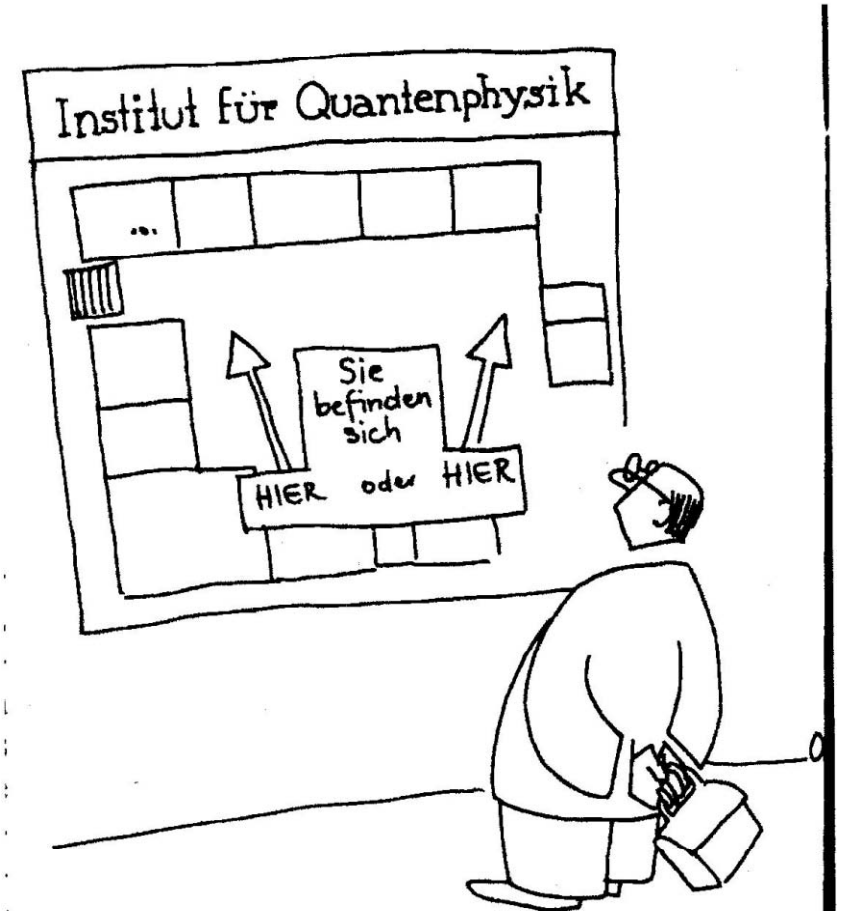


novel 3D structure

Statisticians vs. Physicists

"Data don't make any sense,
we will have to resort to statistics."

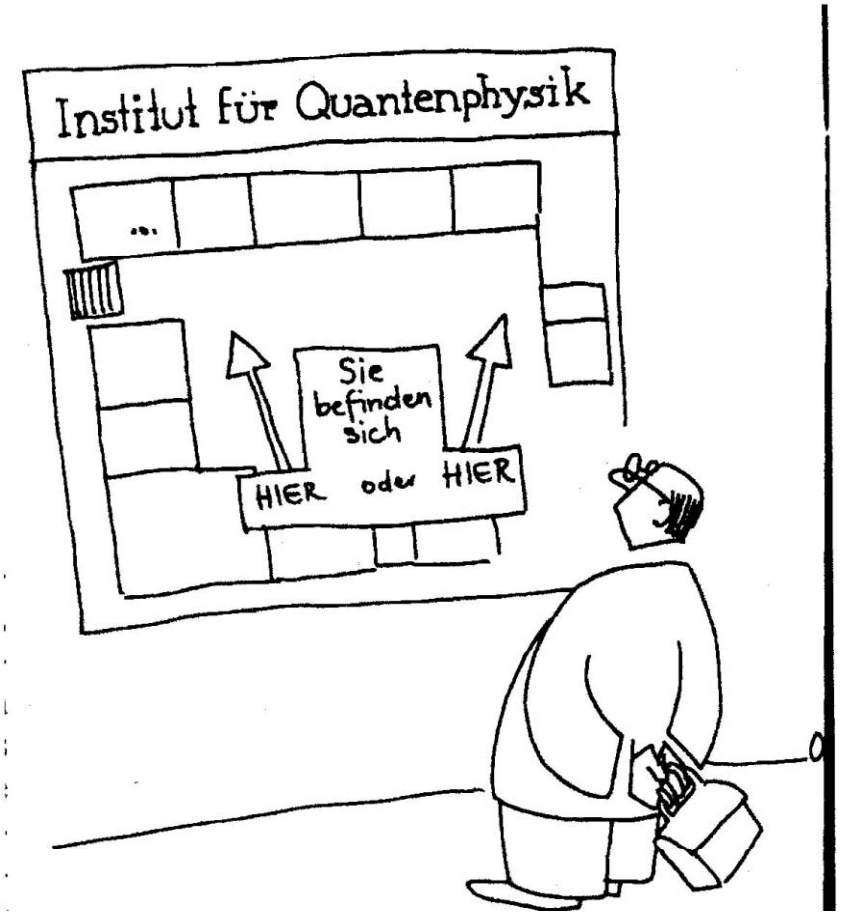
Courtesy of VADLO.com. Used with permission.



What were the key simplifications of the statistical approach?

"Data don't make any sense,
we will have to resort to statistics."

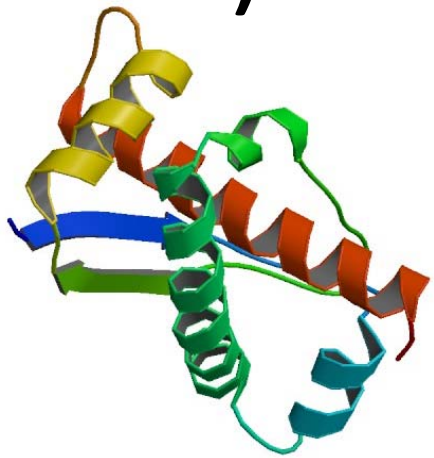
Courtesy of VADLO.com. Used with permission.



Threading (fold recognition)

IQVFLSARPPAPEVSKIY
DNLILQYSPSKSLQMILR
RALGDFENMLADGSFR
AAPKSYPIPHTAFEKSIIV
QTSRMFPVSLIEAARNH
FDPLGLETARAFGHKLA
TAALACFFAREKATNS

?



How could we use the potential energy function to recognize the correct fold?

Methods for Refining Structures

1. Energy minimization
2. Molecular dynamics
3. Simulated Annealing

1. Energy Minimization

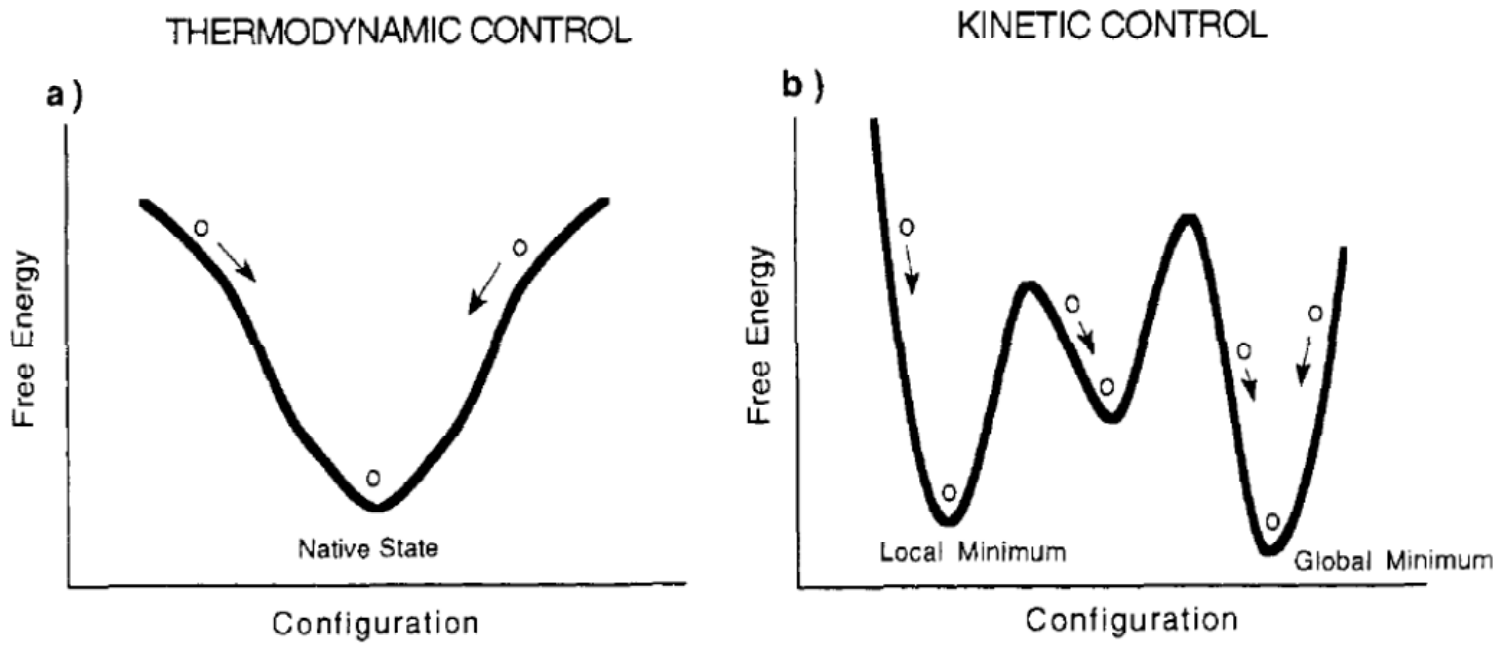
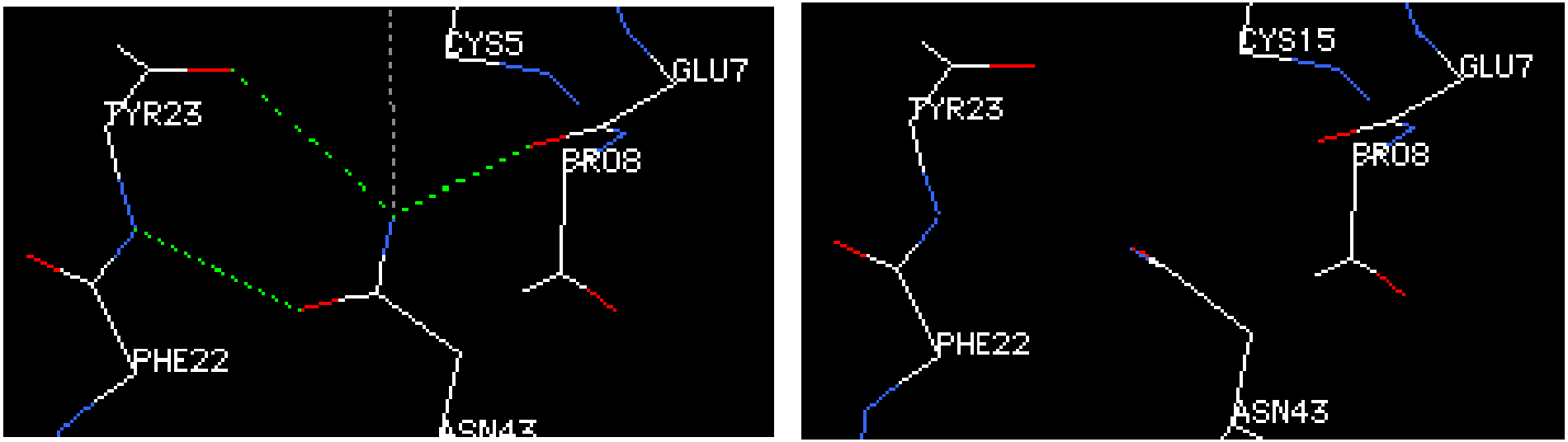


FIGURE 1: Schematic diagram of one-dimensional cross sections through the free energy surfaces of protein folding reactions contrasting two extremes: thermodynamic vs kinetic control. A simple folding surface with a single free energy minimum is shown in panel a. Such a molecule would fold under thermodynamic control, seeking out the most stable state. This is to be contrasted with the considerably more convoluted energy surface in panel b. Because of the high barriers, starting at different locations could lead to different final conformations.

Source: Baker, David, and David A. Agard. "Kinetics Versus Thermodynamics in Protein Folding." *Biochemistry* 33, no. 24 (1994): 7505-9.

Consider a small error in a structure

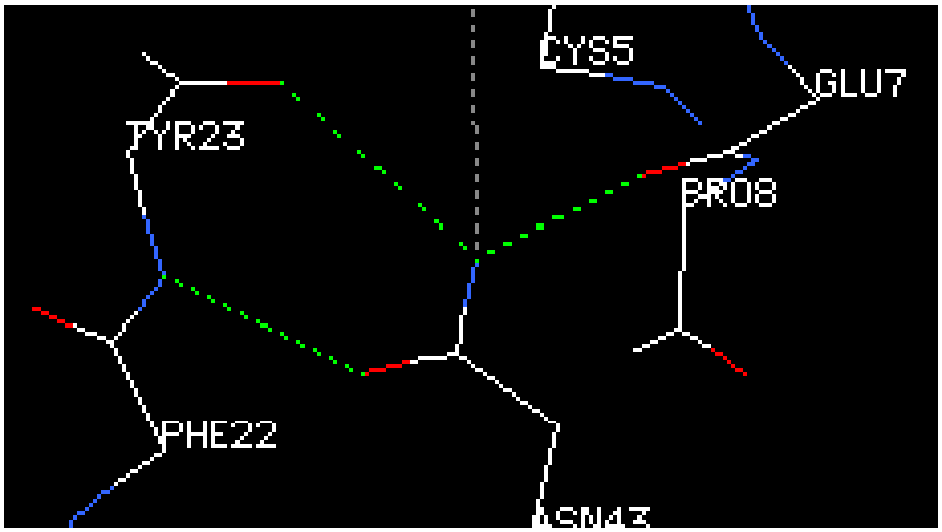
- True structure
- Misplaced side chain

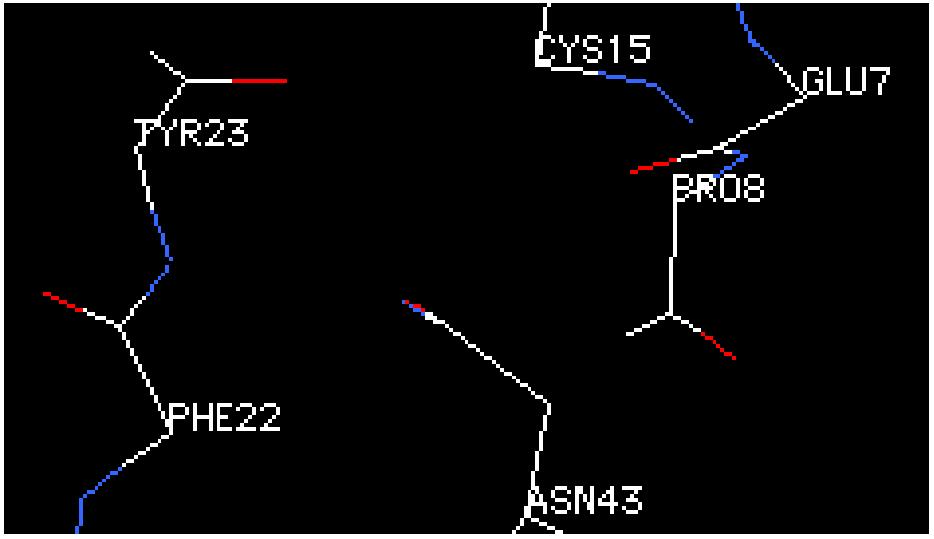


Courtesy of Swiss Institute of Bioinformatics. Used with permission.

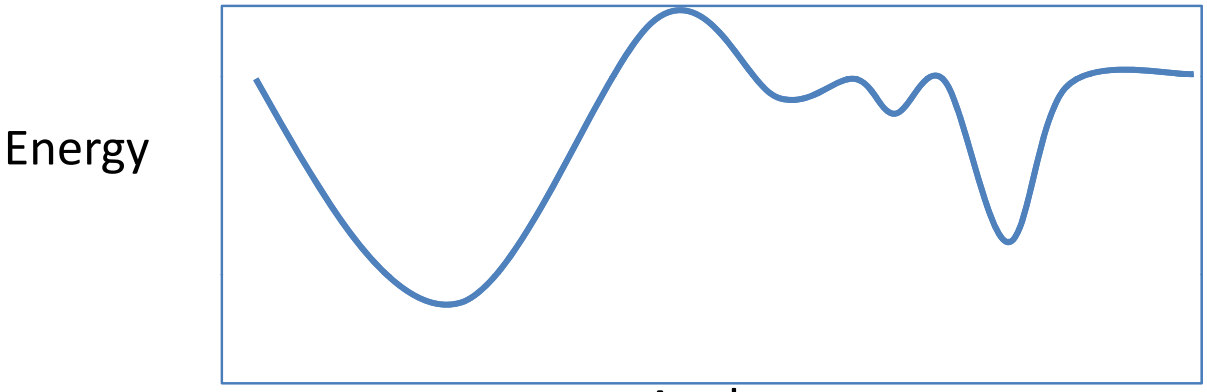
Examples from this good tutorial

Can we restore the side chain?



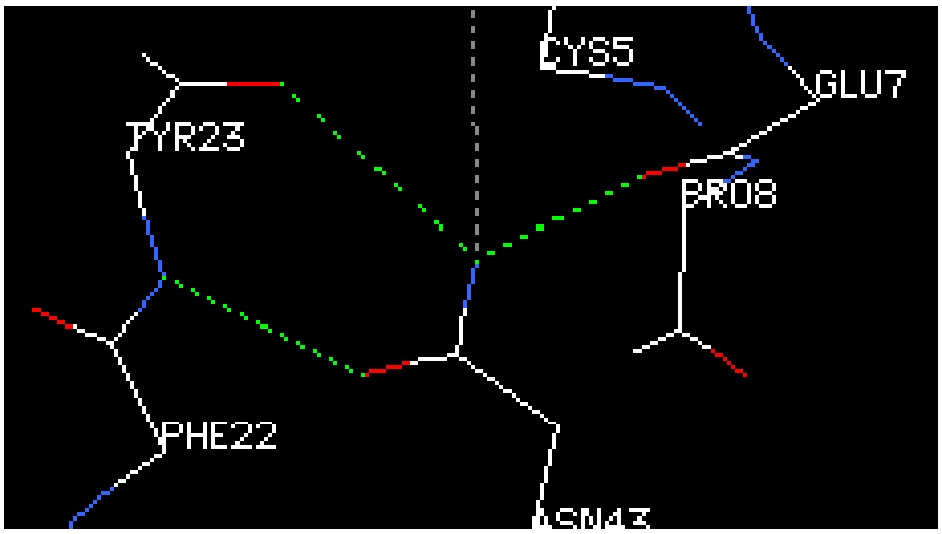


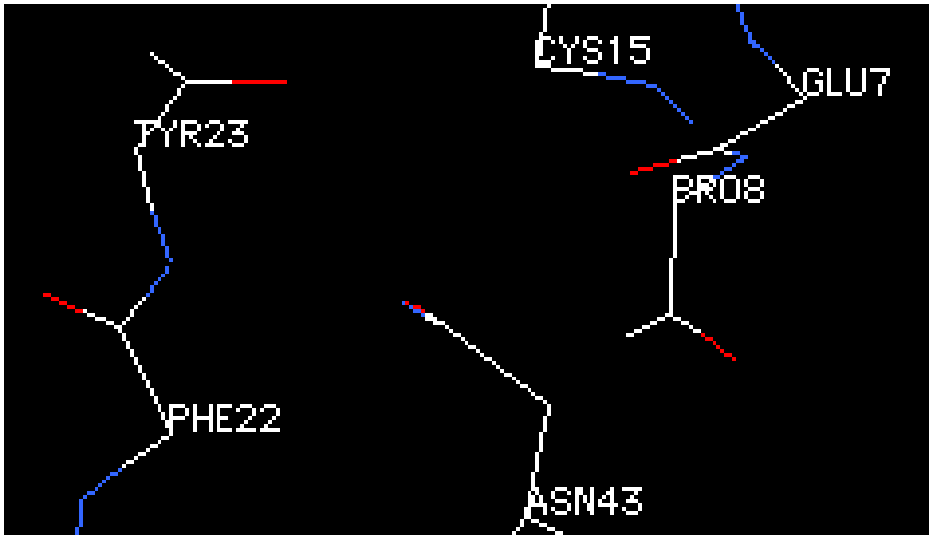
Courtesy of Swiss Institute of Bioinformatics. Used with permission.



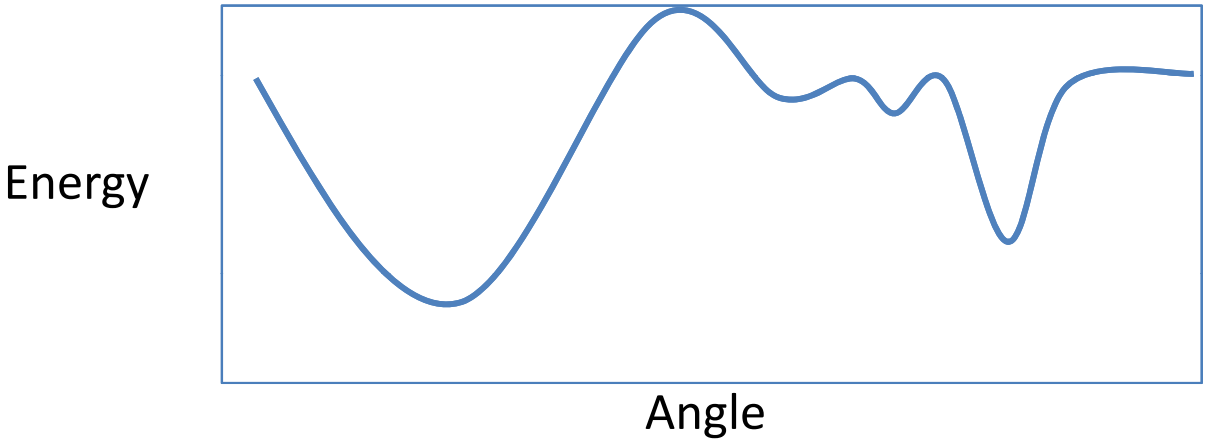
Minimization

- We have equations for U(x,y,z).
- Find nearby values of x,y,z that minimize U.

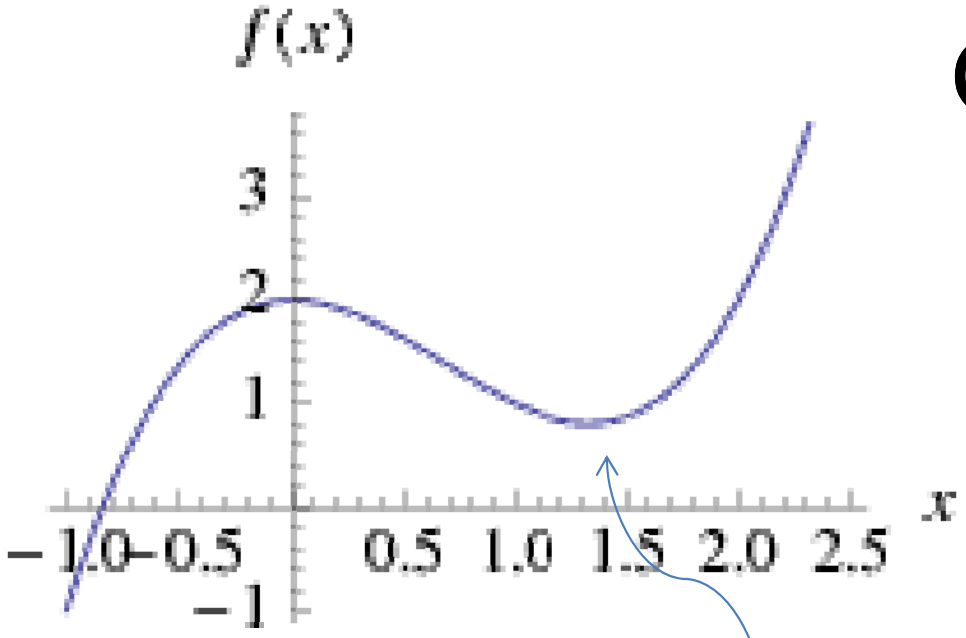




Courtesy of Swiss Institute of Bioinformatics. Used with permission.



$$f(x) = x^3 - 2x^2 + 2$$

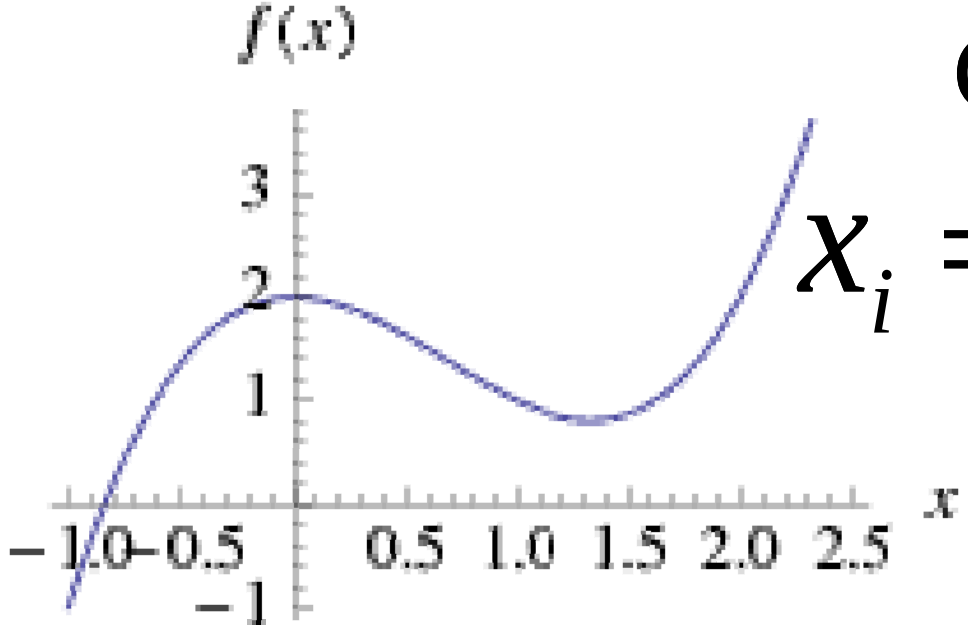


Gradient Descent

$$f'(x) = 0$$

http://mathworld.wolfram.com/MethodofSteepestDescent.html

$$f(x) = x^3 - 2x^2 + 2$$

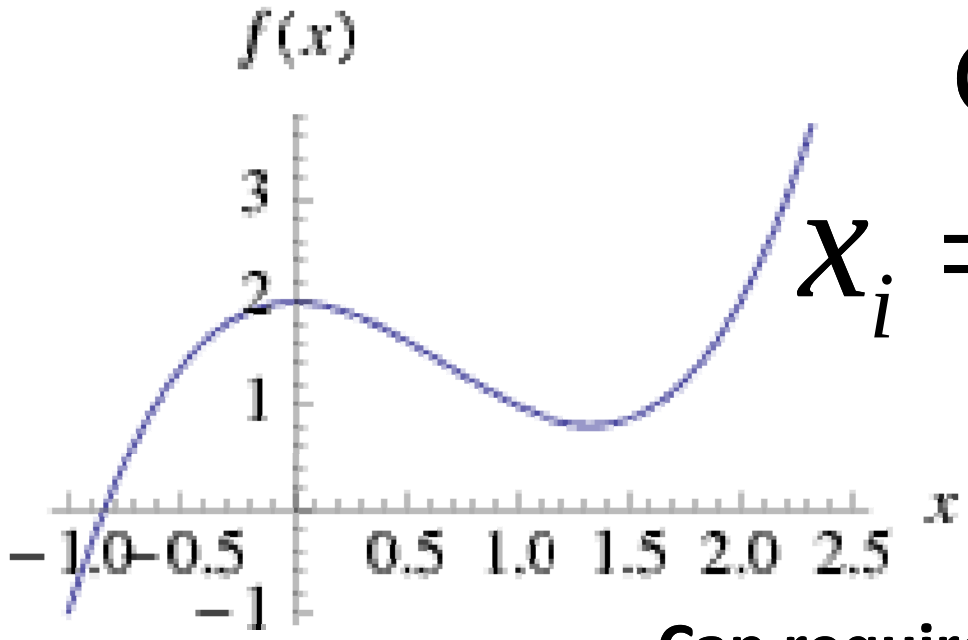


Gradient Descent

$$x_i = x_{i-1} - \varepsilon f'(x_{i-1})$$

http://mathworld.wolfram.com/MethodofSteepestDescent.html

$$f(x) = x^3 - 2x^2 + 2$$

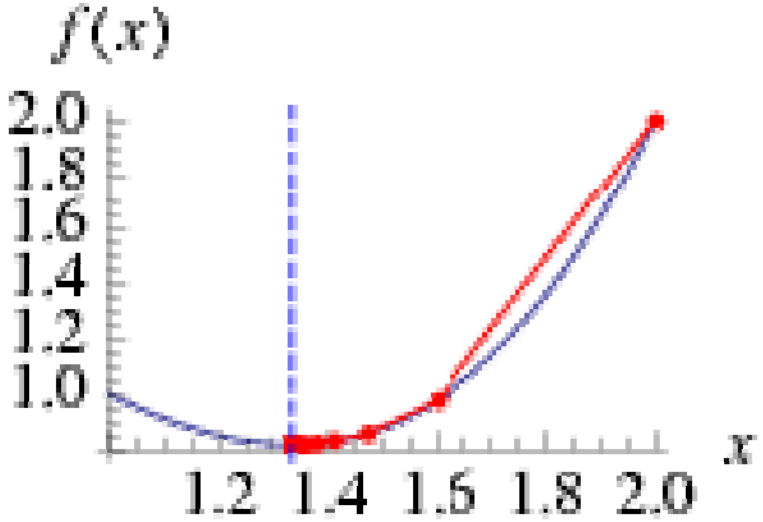


Gradient Descent

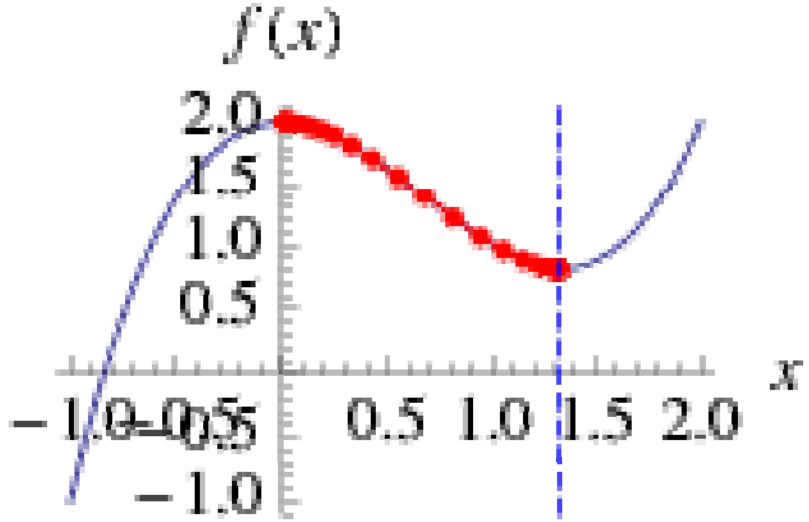
$$x_i = x_{i-1} - \varepsilon f'(x_{i-1})$$

Can require many iterations

$x_0 = 2$



$x_0 = 0.01$



http://mathworld.wolfram.com/MethodofSteepestDescent.html

One dimension $$x_i = x_{i-1} - \varepsilon f'(x_{i-1})$$

N dimensions $$\vec{x}_1 = \vec{x}_0 - \varepsilon \nabla U_0(\vec{x}_0)$$

Where gradient is defined as $$\nabla U = \left(\frac{\partial U}{\partial x_1}, \ldots, \frac{\partial U}{\partial x_n} \right)$$

Since Force is $F = -\nabla U$

$$\vec{x}_1 = \vec{x}_0 + \varepsilon F_0(\vec{x}_0)$$

each step is moving in the direction of the force

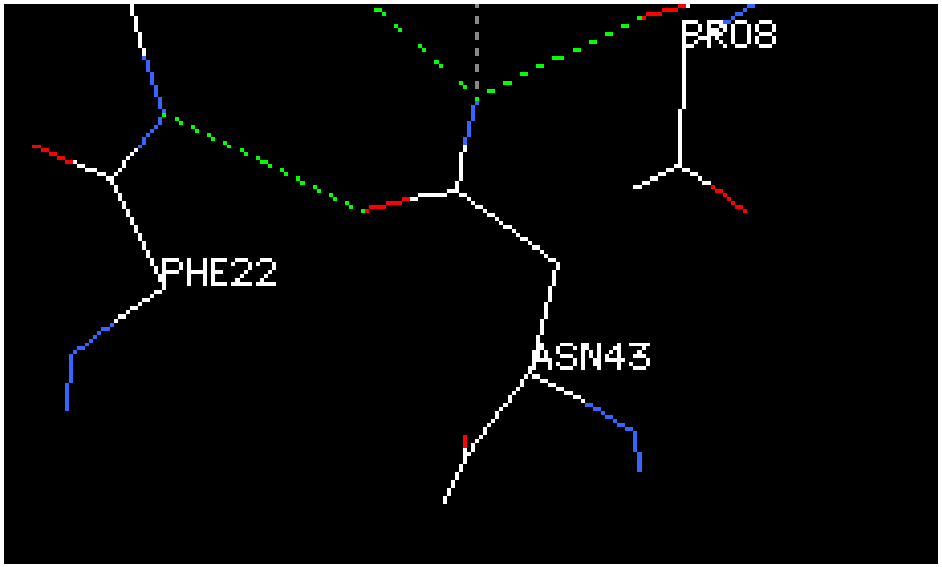
Minimization

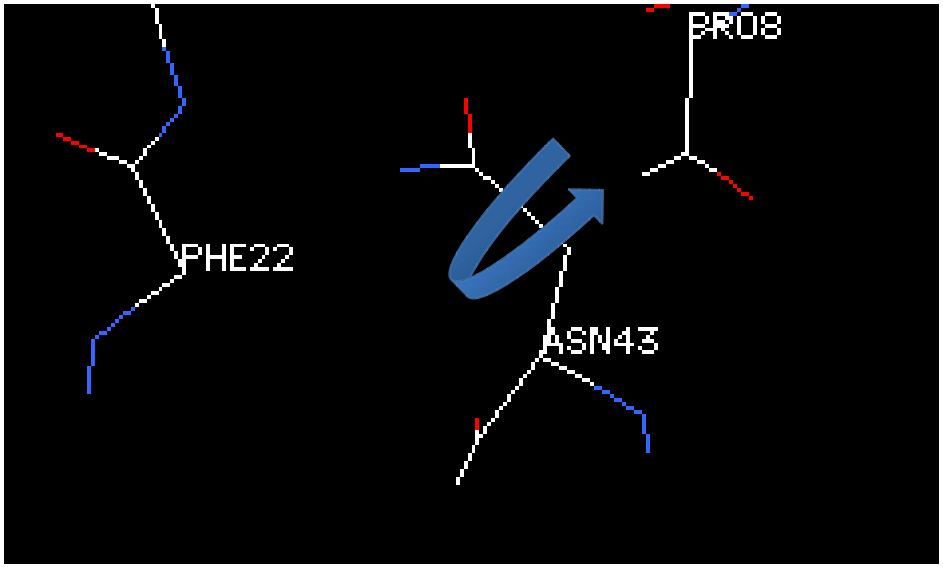
- Convergence can be a problem on some surfaces. More sophisticated approaches are available
- Our example used a continuous energy function, but can be define for discrete optimization too.

Can we restore the side chain?

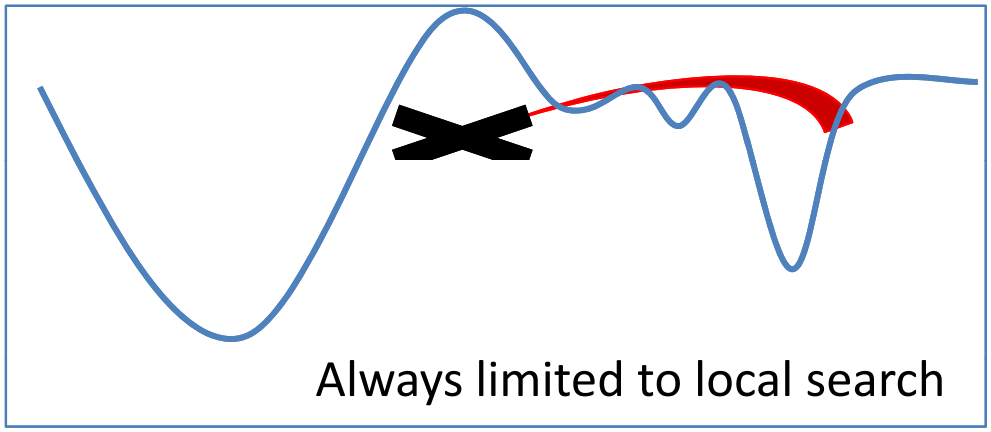
Starting conformation

Unsuccessful minimization





Courtesy of Swiss Institute of Bioinformatics. Used with permission.



Good tutorial

2. Molecular Dynamics

- Seeks to simulate the motion of molecules
- Can escape local minima

$$x(t_i) = x(t_{i-1}) + v(t_{i-1}) \times (t_i - t_{i-1})$$

$$v(t_i) = v(t_{i-1}) + \frac{F(t_{i-1})}{m} \times (t_i - t_{i-1})$$

$$v(t_i) = v(t_{i-1}) - \frac{\nabla U(t_{i-1})}{m} \times (t_i - t_{i-1})$$

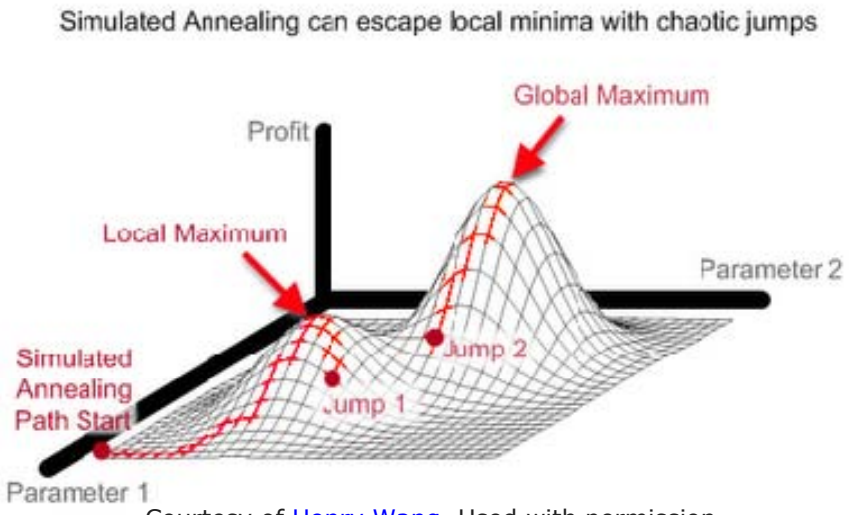
Movie: Simulation of protein folding

Notes

- Short simulations take tremendous computing resources
- Length of simulation and protocol determine **radius of convergence**

3. Simulated Annealing

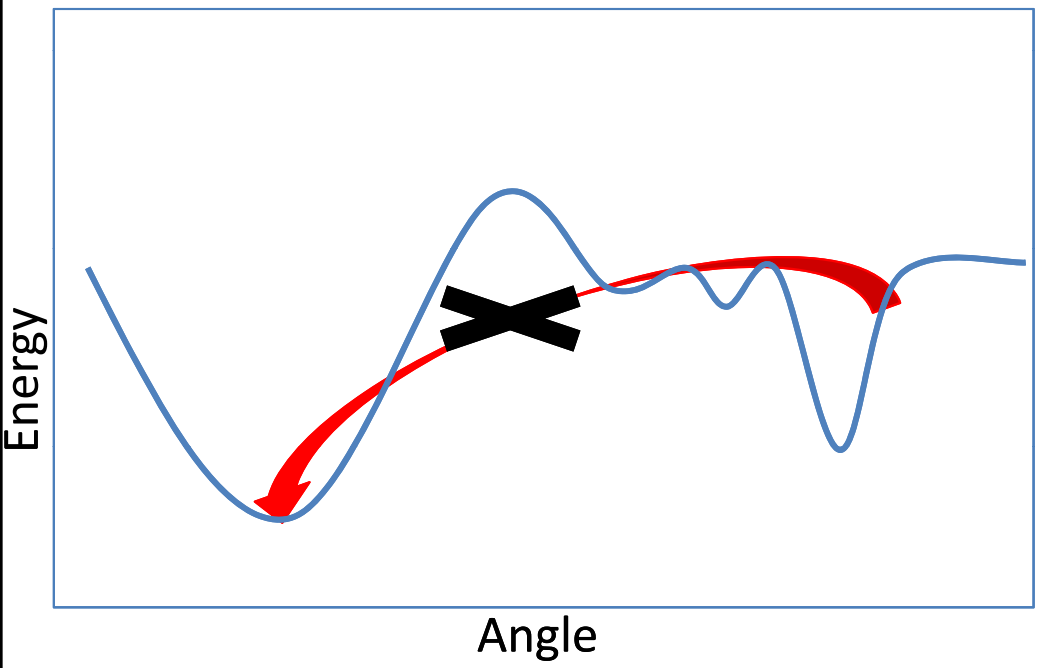
- Physical annealing - high temperature is used to avoid metal defects (local minima).
- Simulated annealing is analogous, and can be applied to many optimization problems

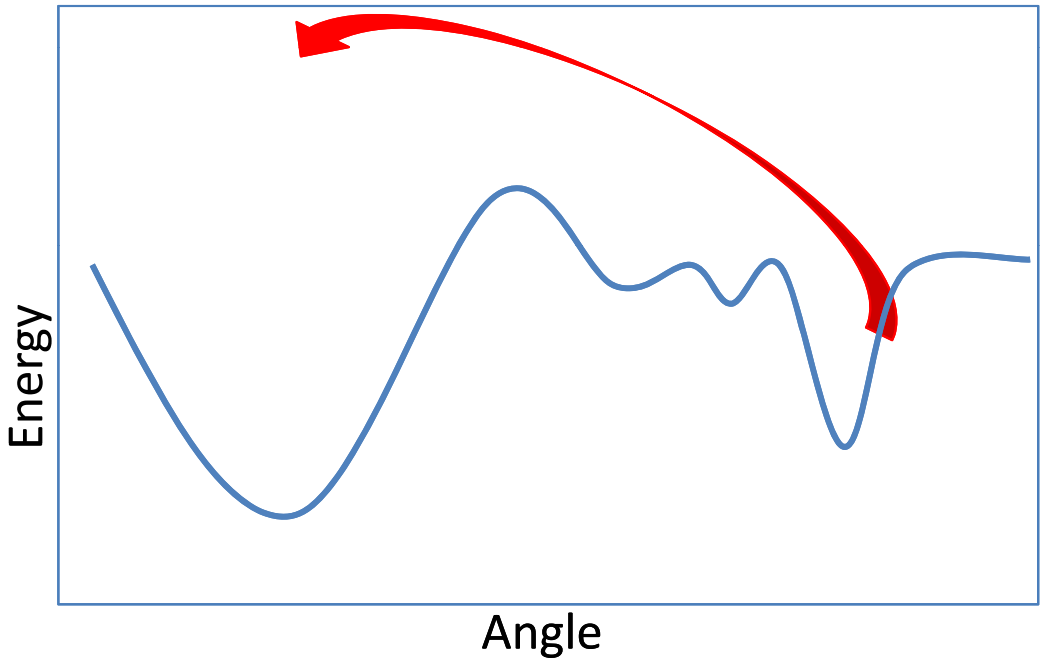


Courtesy of Henry Wang. Used with permission.

3. Simulated Annealing

- At low temperature we cannot escape local minima
- At high temperature the kinetic energy exceeds the potential energy barrier

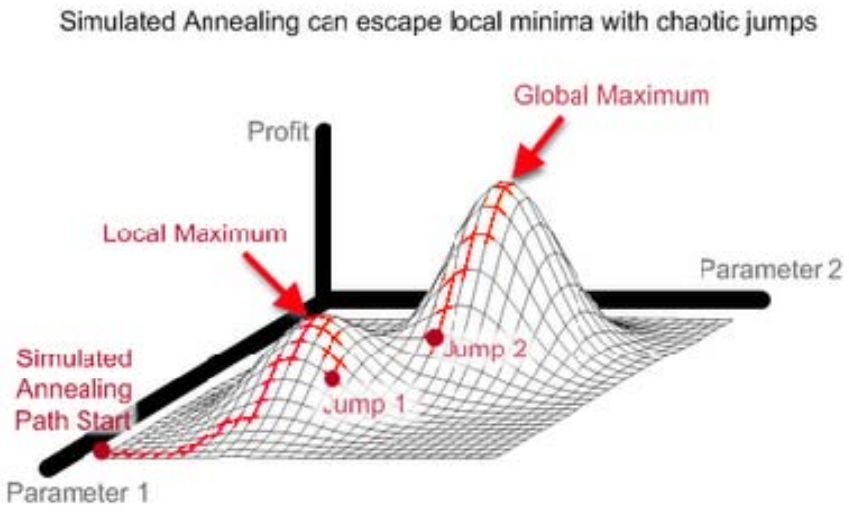




3. Simulated Annealing

- Atoms find equilibrium distribution at higher temperatures

- How can we find the equilibrium distribution of a complicated potential function?



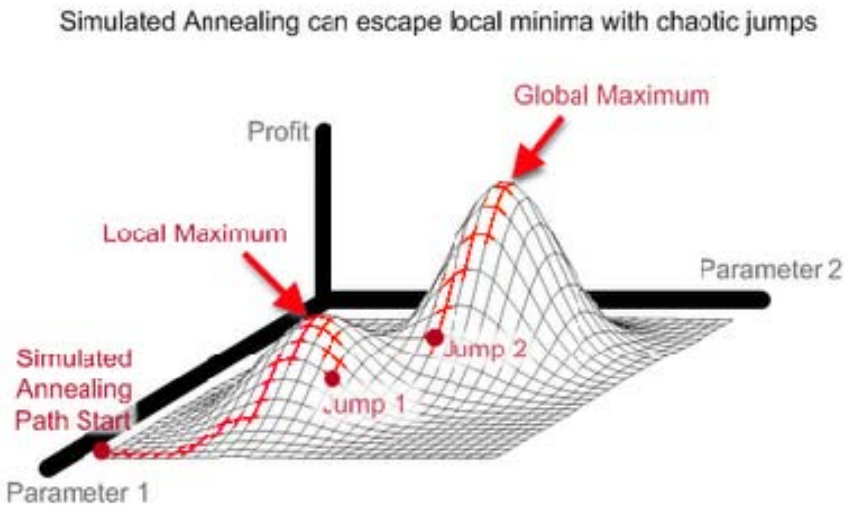
Courtesy of Henry Wang. Used with permission.

http://homesteadingsurvival.myshopify.com/products/115-blacksmithing-forging-welding-metallurgy-sword-books-on-dvd-rom

http://www.stanford.edu/~hwang41/mcmc.png

3. Simulated Annealing

- Start at high temperature
- Find most probable states
- Reduce temperature to trap these states



Courtesy of Henry Wang. Used with permission.

http://homesteadingsurvival.myshopify.com/products/115-blacksmithing-forging-welding-metallurgy-sword-books-on-dvd-rom

http://www.stanford.edu/~hwang41/mcmc.png

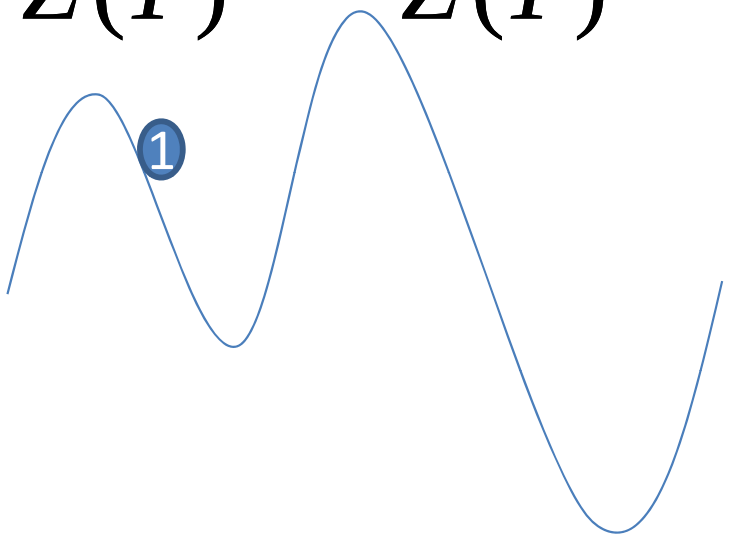
Metropolis Algorithm

- Goal: efficiently search a large conformation space.
- Can be understood in terms of physical processes, but much more general
- Note the difference from molecular dynamics:
  - Molecules move under physical forces but temperatures are far outside of normal range
  - **A sampling method not a simulation**!

Acceptance Criteria

- Randomly choose neighboring state:
  - Always accept moves that reduce potential
  - Go uphill (higher potential) based on odds ratio

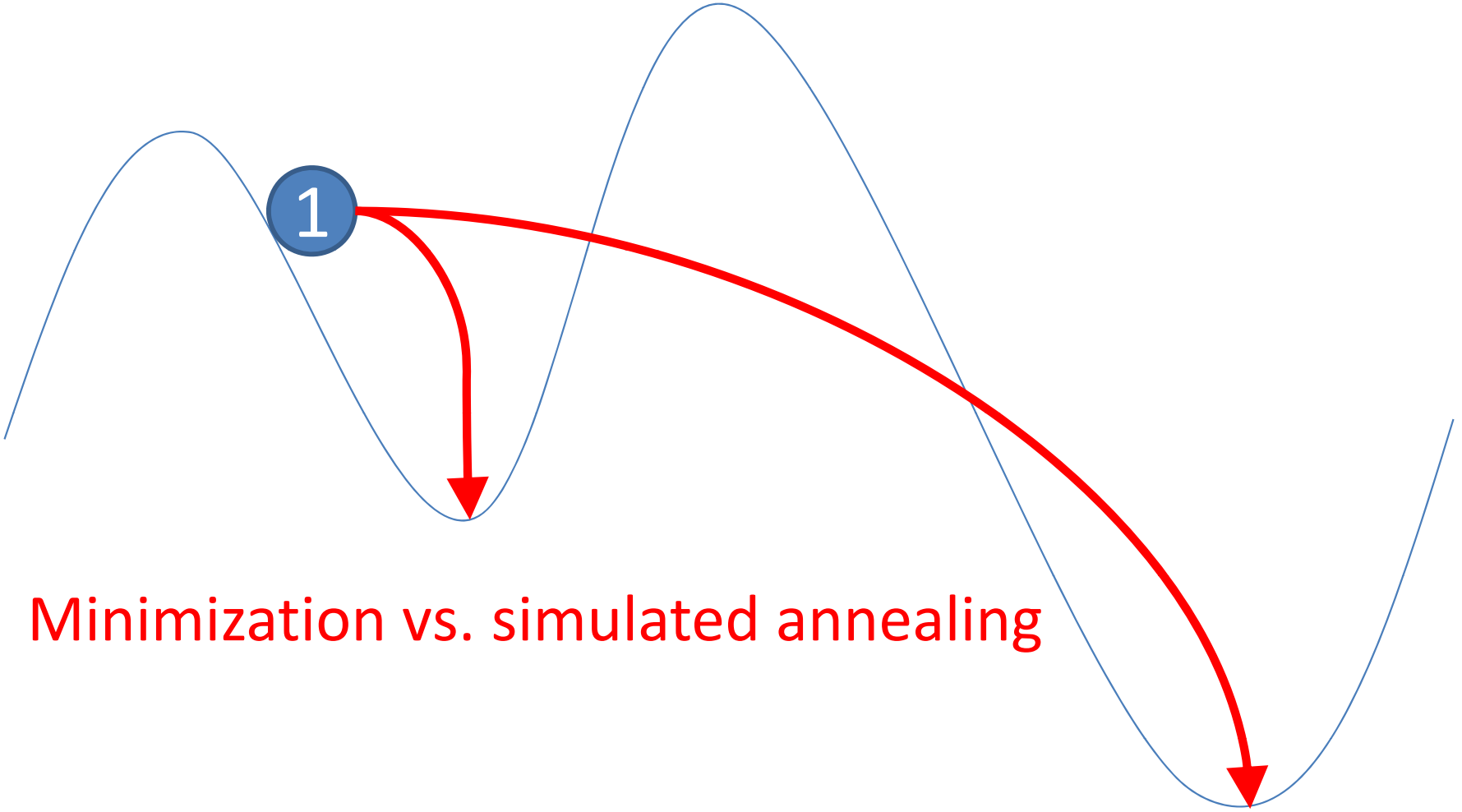
$$\frac{P(Stest)}{P(Sn)} = \frac{e^{-Etest/kT}}{Z(T)} / \frac{e^{-En/kT}}{Z(T)} = e^{-(Etest - E_n)/kT}$$



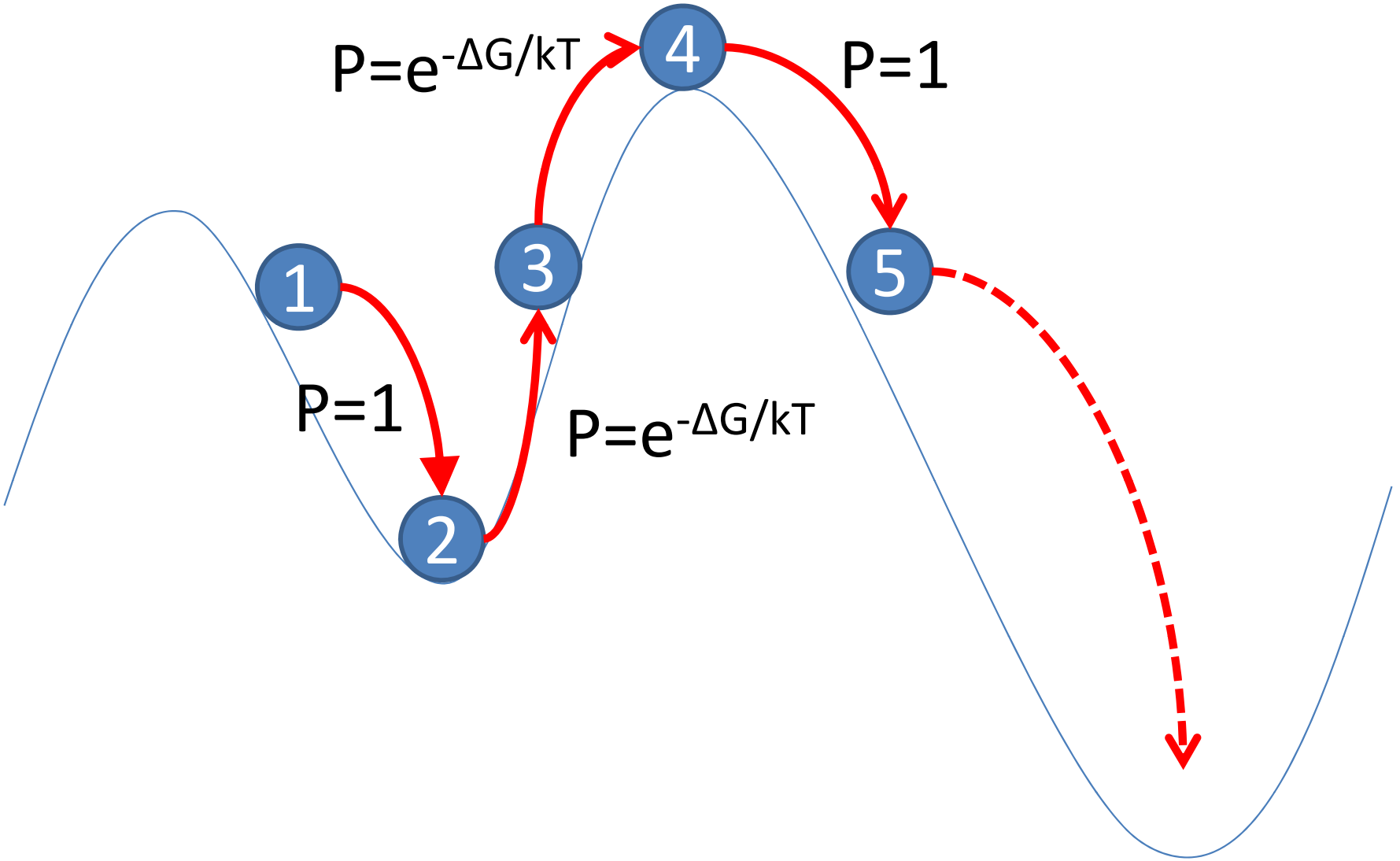
3. Metropolis sampling

Iterate for a fixed number of cycles or until convergence:

1. Start with a system in state S_n with energy E_n
2. Choose a neighboring state at random; we will call it the proposed state : S_{test} with energy E_{test}
3. If $E_{test} < E_n$: $S_{n+1} = S_{test}$
4. Else set $S_{n+1} = S_{test}$ with probability $P = e^{-(Etest - E_n)/kT}$
   - otherwise $S_{n+1} = S_n$



Minimization vs. simulated annealing


P=e-ΔG/kT
4
P=1
1
3
5
P=1
P=e-ΔG/kT
2

Acceptance Criteria

- Always go down-hill
- Go uphill based on odds ratio

$$\frac{P(Stest)}{P(Sn)} = \frac{e^{-Etest/kT}}{Z(T)} / \frac{e^{-En/kT}}{Z(T)} = e^{-(Etest-E_n)/kT}$$

How does T alter outcome?

Acceptance Criteria

- Always go down-hill
- Go uphill based on odds ratio

$$\frac{P(Stest)}{P(Sn)} = \frac{e^{-Etest/kT}}{Z(T)} / \frac{e^{-En/kT}}{Z(T)} = e^{-(Etest - E_n)/kT}$$

Annealing schedule:

- Start at High T
- Lower slowly

3. Metropolis sampling – prob. version

To identify minima given a probability function: P(S)

1. Start with a system in state S_n
2. Choose a neighboring state at random : S_{test}
3. Compute acceptance ratio $a = \frac{P(S_{test})}{P(S_N)}$
4. If a> 1 : S_{n+1}=S_{test}
5. Else set S_{n+1}= S_{test} with probability *a* and S_{n+1}= S_n with probability *1-a*

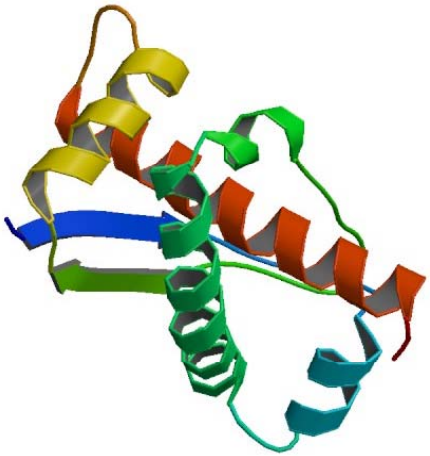
Not specific to protein structure. Used to sample diverse probability distributions

Review:
Methods for Refining Structures

1. Energy minimization
2. Molecular dynamics
3. Simulated annealing

Methods for Predicting Structure

IQVFLSARPPAPEVSKIY
DNLILQYSPSKSLQMILR
RALGDFENMLADGSFR
AAPKSYPIPHTAFEKSIIV
QTSRMFPVSLIEAARNH
FDPLGLETARAFGHKLA
TAALACFFAREKATNS



novel 3D structure

What actually works for structure prediction?

CASP1

(First meeting on Critical Assessment of techniques for protein Structure Prediction)

A Large-Scale Experiment to Assess Protein Structure Prediction Methods

PROTEINS: Structure, Function, and Genetics 23:ii–iv (1995)

COLLECTING PREDICTION TARGETS

Information was solicited from X-ray crystallographers and NMR spectroscopists about structures that were either expected to be solved shortly or that had been solved already but not discussed in public. Targets were identified through personal contacts, blanket emailing, and appeals at scientific meetings. The collecting and management of prediction targets proved to be a difficult undertaking. In all, information on 33 different proteins was obtained. Some of these were not solved in time for the prediction experiment and some were made public without sufficient notice to the predictors. Finally, one or more predictions were received on 24 of these targets.

Rosetta

Raman et al. Proteins 2009; 77(Suppl 9):89–99.

http://onlinelibrary.wiley.com/doi/10.1002/prot.22540/full

- Two types of models:
  - –Homology
  - –*de novo*

Homology

- Align query to sequences in PDB
- Use several alignment methods
- Three categories of queries:
  1. High sequence similarity template(s) (>50% sequence similarity).
  2. Medium sequence similarity template(s) (20–50% sequence similarity).
  3. Low sequence similarity template(s) (<20% sequence similarity).

Homology

- Align query to sequences in PDB
- Use several alignment methods
- Refine models

tools you have seen earlier in the course

General Refinement Procedure

- Random changes to backbone torsion angles
- Rotamer optimization of side chains
- Energy minimization of torsion angles (bond lengths and angles kept fixed)

Homology

High sequence similarity template(s) (>50% sequence similarity).

– Minimal refinement, focused on regions where alignment is poor.

Homology

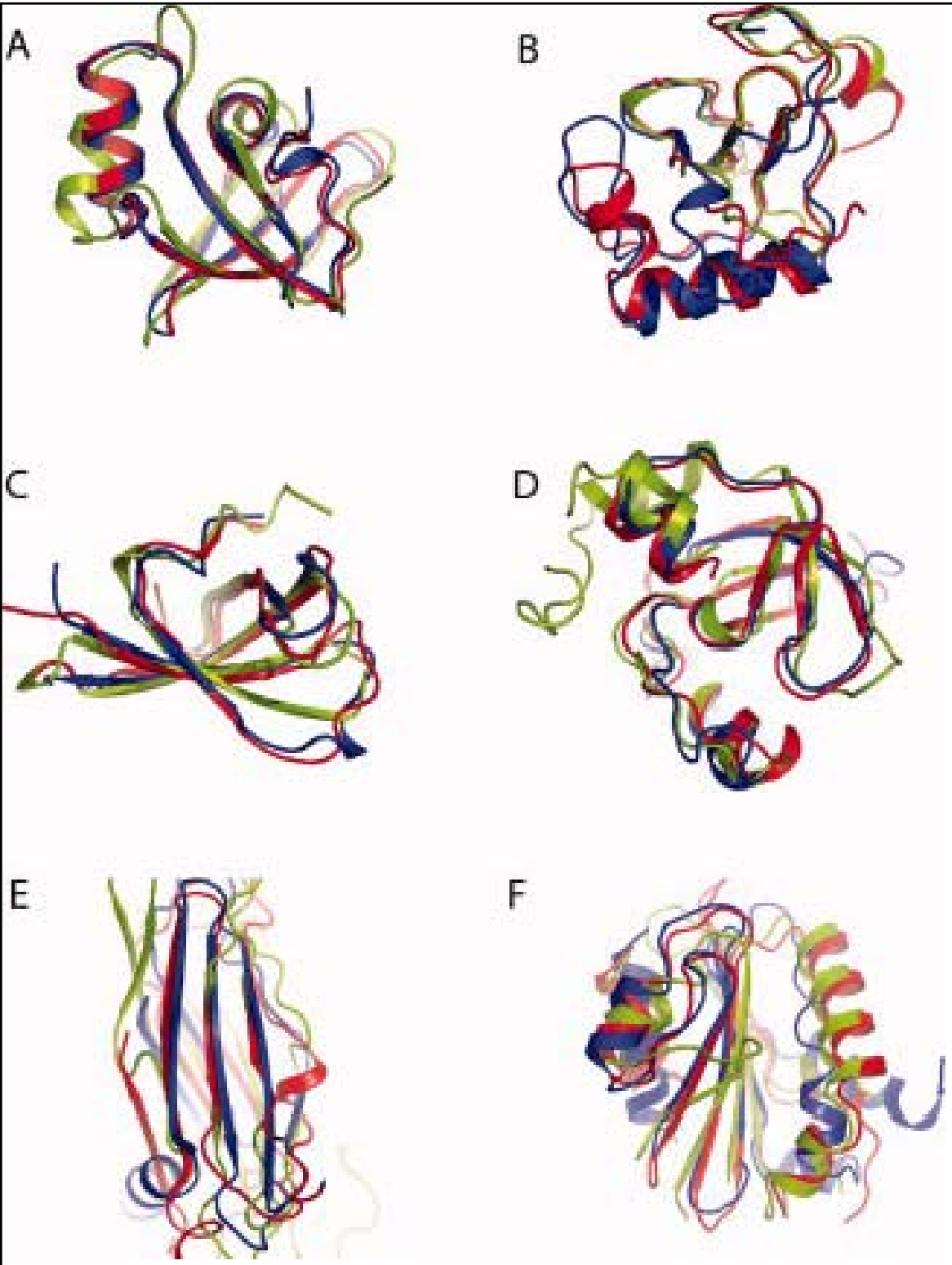
Medium sequence similarity template(s) (20–50% sequence similarity).

- Proceed with several alignments
- Refine structures
- Choose best model by final energy

Homology

Medium sequence similarity template(s) (20–50% sequence similarity).

- Refinement focuses on regions near gaps and insertions, loops in the starting model, and sequence segments with low conservation
- Replaces torsion angles with those from peptides of known structure
- Minimize local structure
- Refine global structure



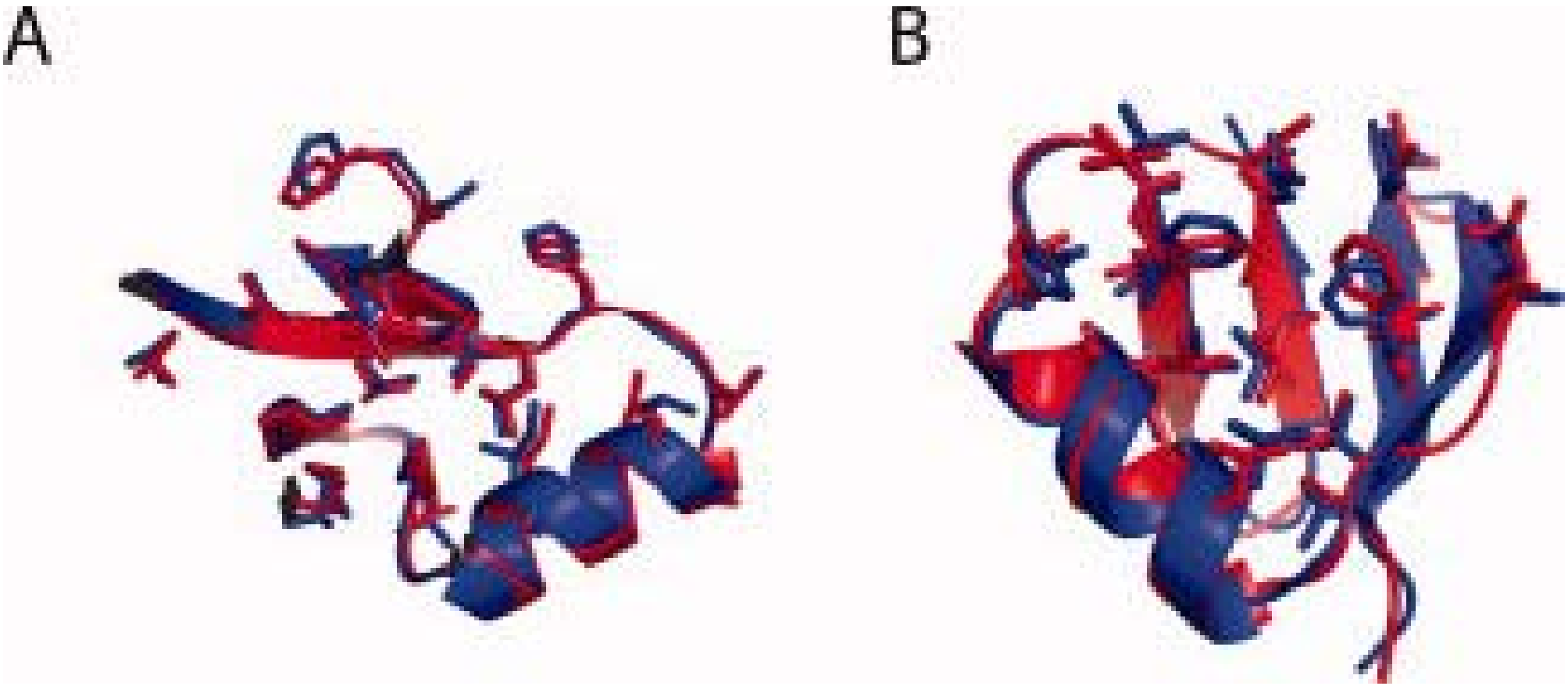
Native structure

Best model

Best template

Source: Raman, Srivatsan, Robert Vernon, et al. "Structure Prediction for CASP8 with All-atom Refinement using Rosetta." *Proteins: Structure, Function, and Bioinformatics* 77, no. S9 (2009): 89-99.

Accurate side chains in core

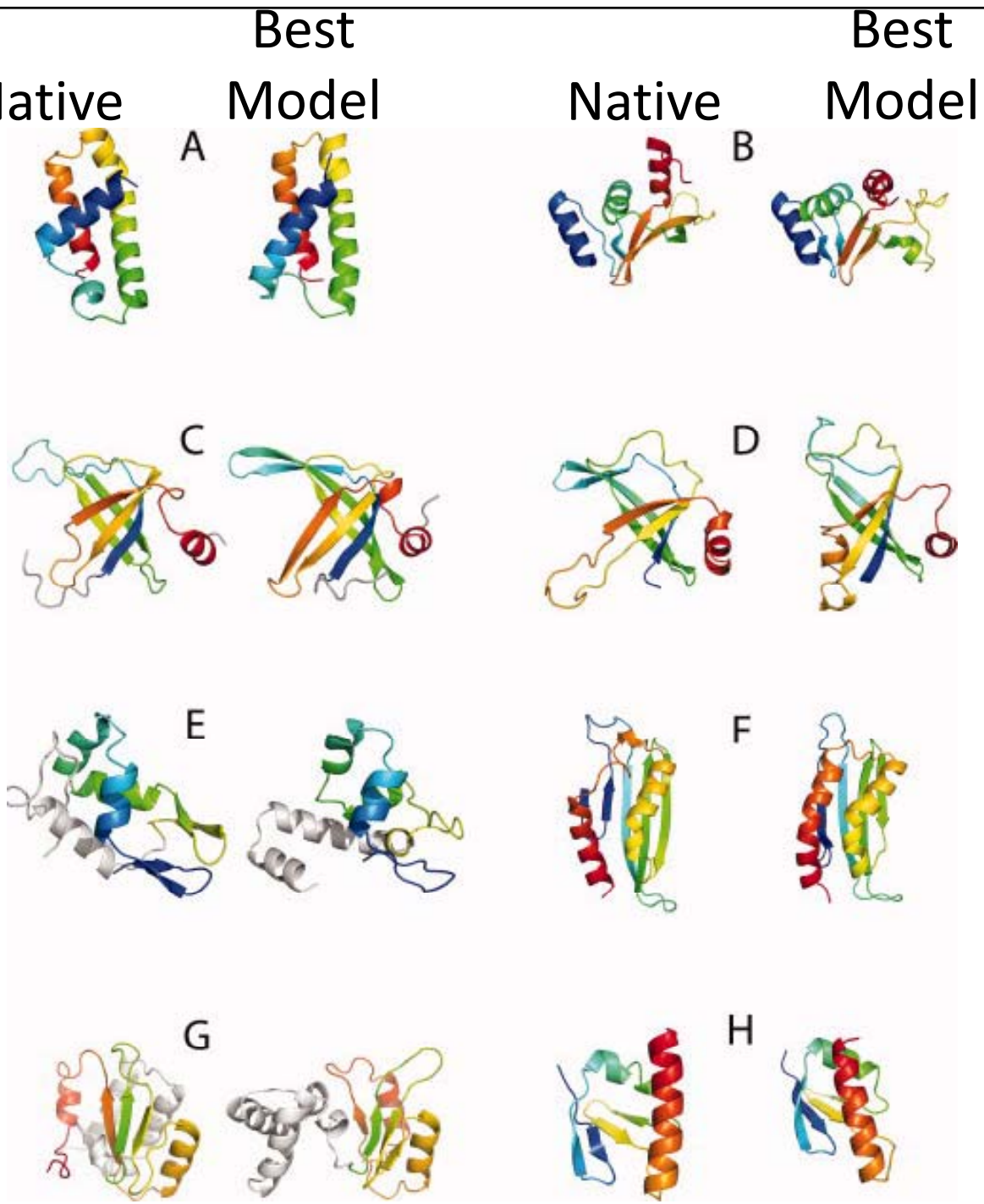


Source: Raman, Srivatsan, Robert Vernon, et al. "Structure Prediction for CASP8 with All-atom Refinement using Rosetta." *Proteins: Structure, Function, and Bioinformatics* 77, no. S9 (2009): 89-99.

Homology

Low sequence similarity template(s) (<20% sequence similarity).

- Use many more starting models
- More aggressive refinement strategy
  - Rebuild secondary structure elements in addition to regions refined in medium homology:
    - gaps and insertions
    - loops in the starting model
    - regions with low conservation



Source: Raman, Srivatsan, Robert Vernon, et al. "Structure Prediction for CASP8 with All-atom Refinement using Rosetta." *Proteins: Structure, Function, and Bioinformatics* 77, no. S9 (2009): 89-99.

de novo

- When there is no suitable homology model:
  - Monte Carlo search for backbone angles
    - Choose a short region (3-9 amino acids) of backbone
    - Set torsion angles to those of a similar peptide in PDB
    - Accept with metropolis criteria
  - 36,000 MC steps
  - Repeat entire process to get 2,000 final structures
  - Cluster structures
  - Refine clusters

How has modeling changed during the CASP challenges?

CASP1

(First meeting on Critical Assessment of techniques for protein Structure Prediction)

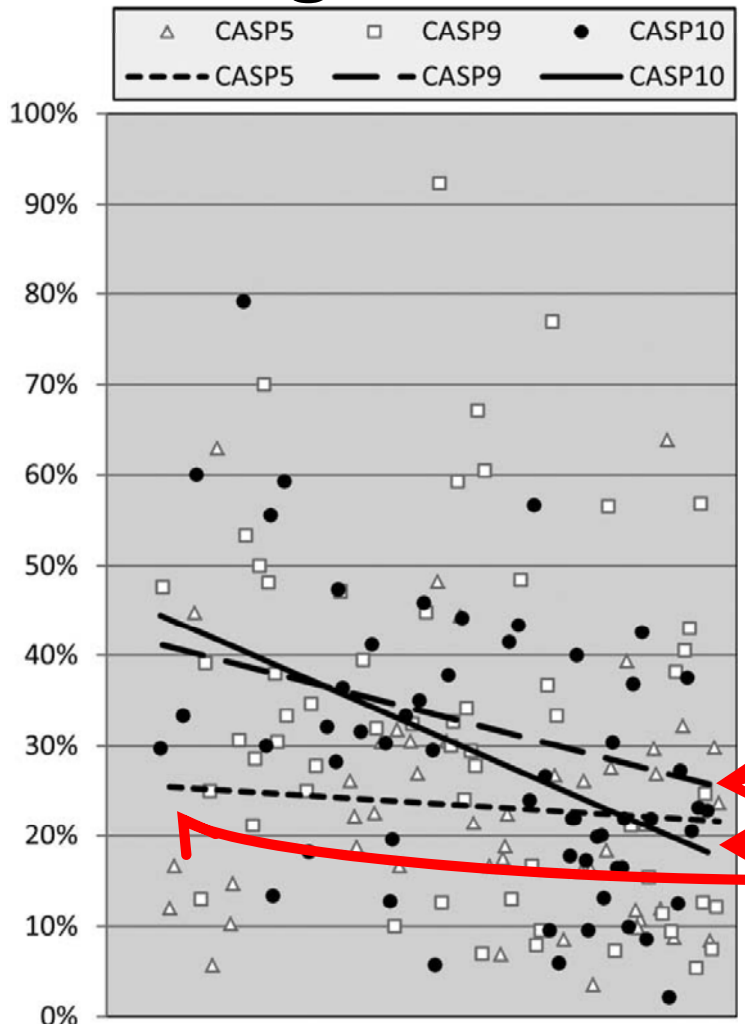
A Large-Scale Experiment to Assess Protein Structure Prediction Methods

PROTEINS: Structure, Function, and Genetics 23:ii–iv (1995)

COLLECTING PREDICTION TARGETS

Information was solicited from X-ray crystallographers and NMR spectroscopists about structures that were either expected to be solved shortly or that had been solved already but not discussed in public. Targets were identified through personal contacts, blanket emailing, and appeals at scientific meetings. The collecting and management of prediction targets proved to be a difficult undertaking. In all, information on 33 different proteins was obtained. Some of these were not solved in time for the prediction experiment and some were made public without sufficient notice to the predictors. Finally, one or more predictions were received on 24 of these targets.

Improvement over the last decade: Percentage of residues successfully modeled



% of residues modeled that were not in best template

CASP10 results compared to those of previous CASP experiments
Andriy Kryshtafovych, Krzysztof Fidelis, John Moult
DOI: 10.1002/prot.24448

Each point represents the best model for a target

CASP10 and CASP9 are similar, much better than CASP5.

Source: Kryshtafovych, Andriy, Krzysztof Fidelis, et al. "CASP10 Results Compared to those of Previous CASP Experiments." *Proteins: Structure, Function, and Bioinformatics* 82, no. S2 (2014): 164-74.

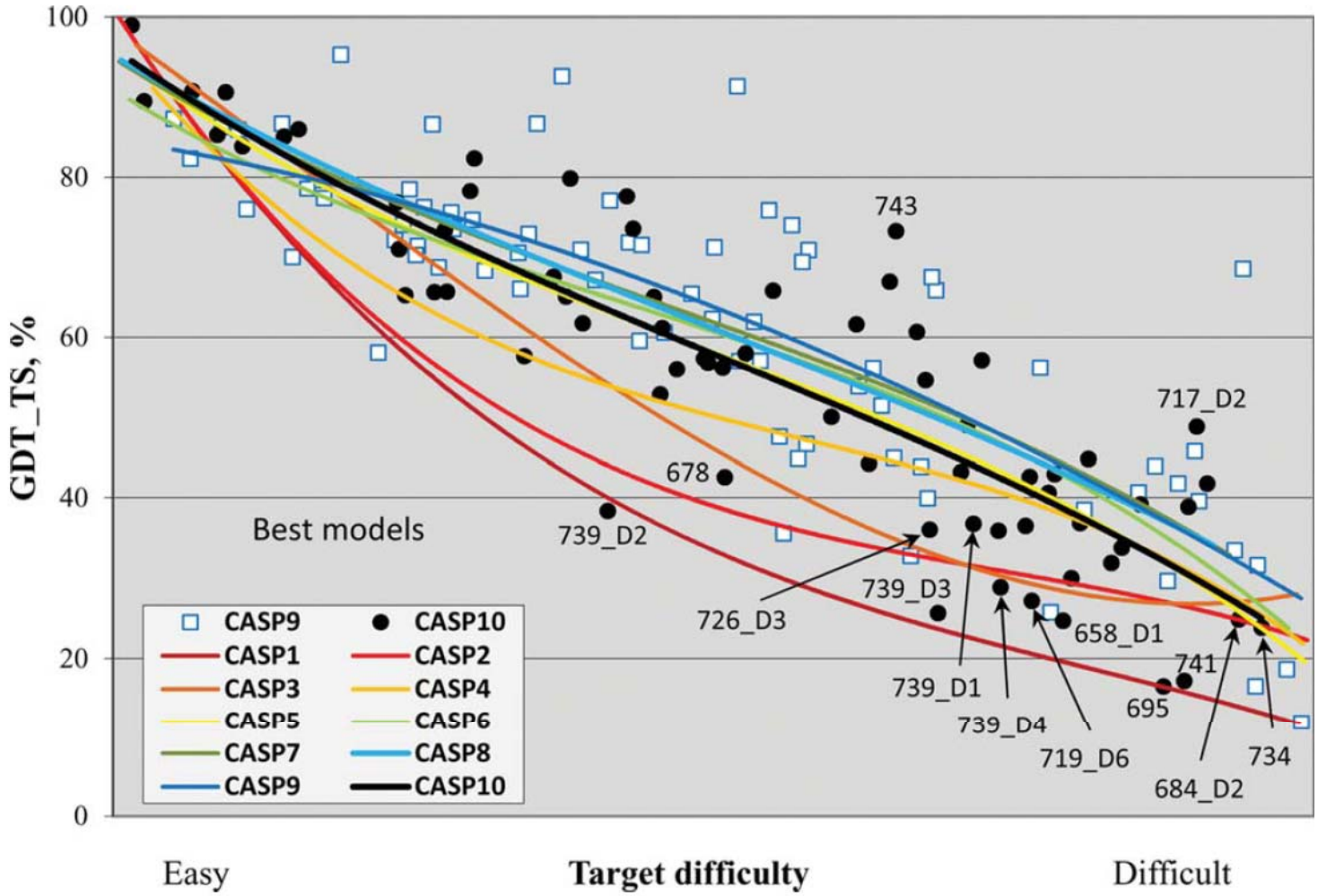
Target difficulty -based on structural and sequence similarity of a target to proteins of known structure

Overall Prediction Accuracy Did Not Improve

CASP10 results compared to those of previous CASP experiments
DOI: 10.1002/prot.24448

Global distance test
GDT_TS
=overall accuracy of a model
average % of Cα atoms in the prediction close to corresponding atoms in the target structure

Perfect model: 90-100
Random model: 20-30



Source: Kryshtafovych, Andriy, Krzysztof Fidelis, et al. "CASP10 Results Compared to those of Previous CASP Experiments." *Proteins: Structure, Function, and Bioinformatics* 82, no. S2 (2014): 164-74.

Target difficulty is based on structural and sequence similarity of a target to proteins of known structure

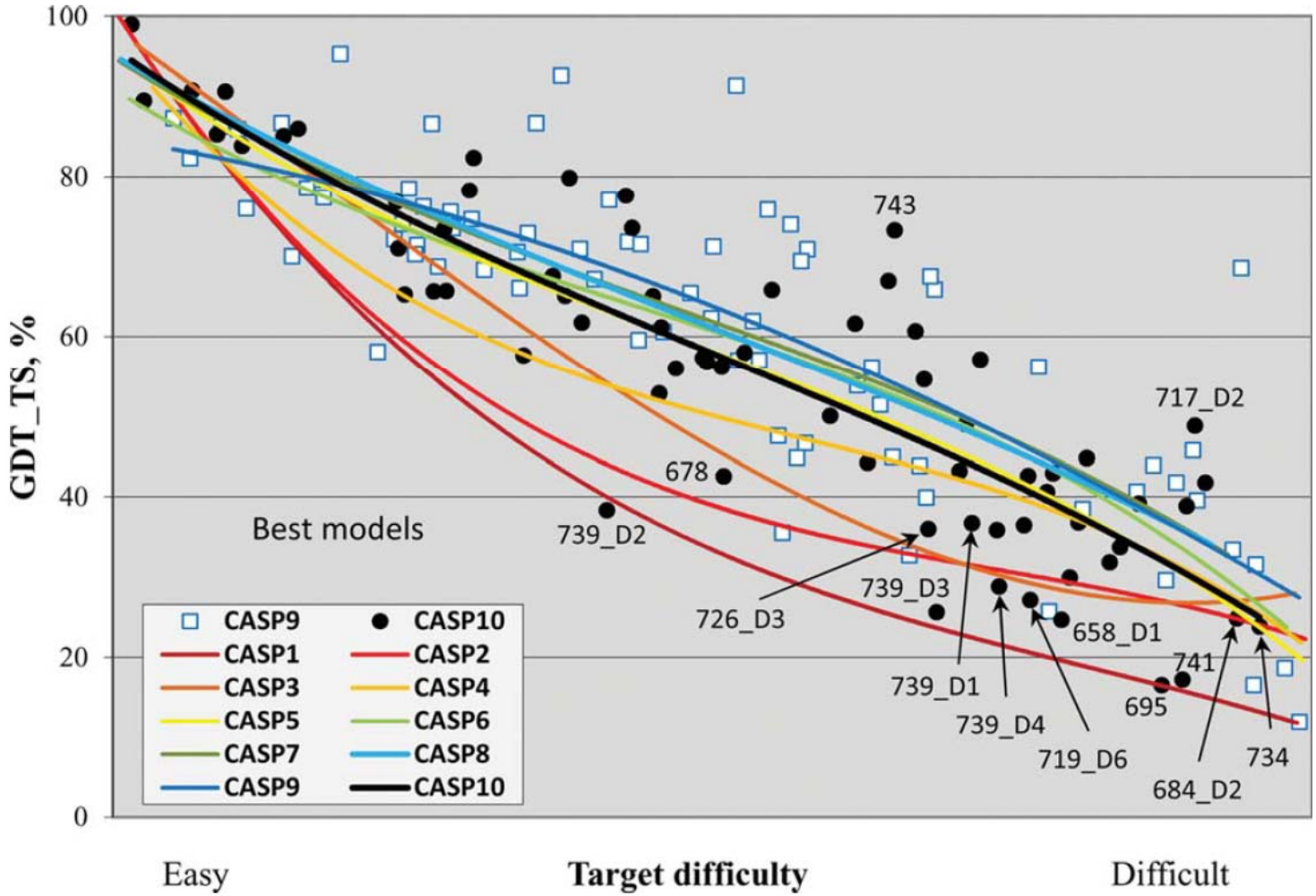
Overall Prediction Accuracy Did Not Improve

CASP10 results compared to those of previous CASP experiments

DOI: 10.1002/prot.24448

Why is the trend line the same?

Are targets getting harder in other ways? Multi-domain, multi-chain, etc.



Source: Kryshtafovych, Andriy, Krzysztof Fidelis, et al. "CASP10 Results Compared to those of Previous CASP Experiments." *Proteins: Structure, Function, and Bioinformatics* 82, no. S2 (2014): 164-74.

Target difficulty is based on structural and sequence similarity of a target to proteins of known structure

Free modeling results

de novo

(no template)

Free Modeling in Flux

CASP10 results compared to those of previous CASP experiments

Andriy Kryshtafovych, Krzysztof Fidelis, John Moult

DOI: 10.1002/prot.24448

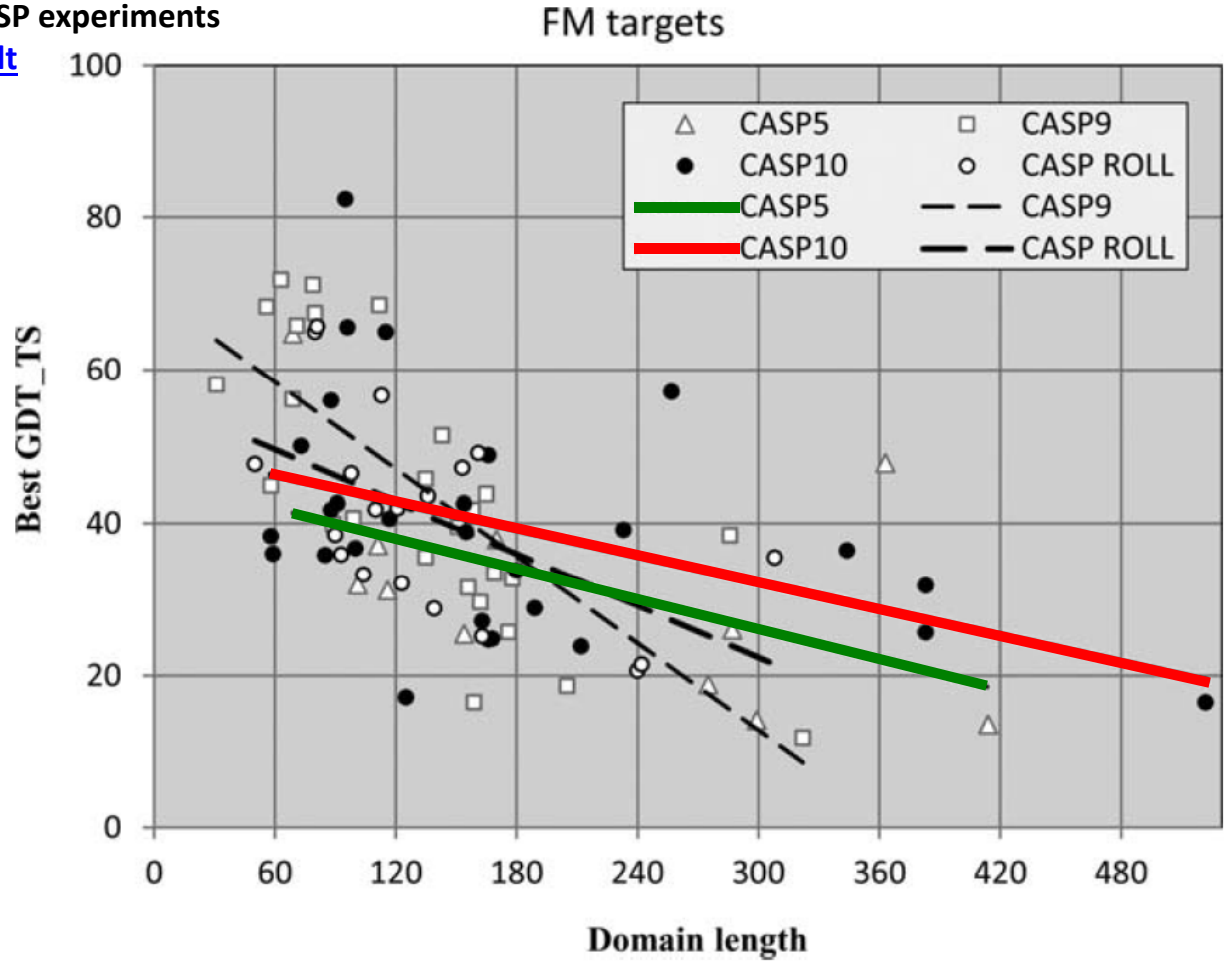
Global distance test

GDT_TS =overall accuracy of a model

average percentage of Cα atoms in the prediction close to corresponding atoms in the target structure

Perfect model: 90-100

Random model: 20-30



Source: Kryshtafovych, Andriy, Krzysztof Fidelis, et al. "CASP10 Results Compared to those of Previous CASP Experiments." *Proteins: Structure, Function, and Bioinformatics* 82, no. S2 (2014): 164-74.

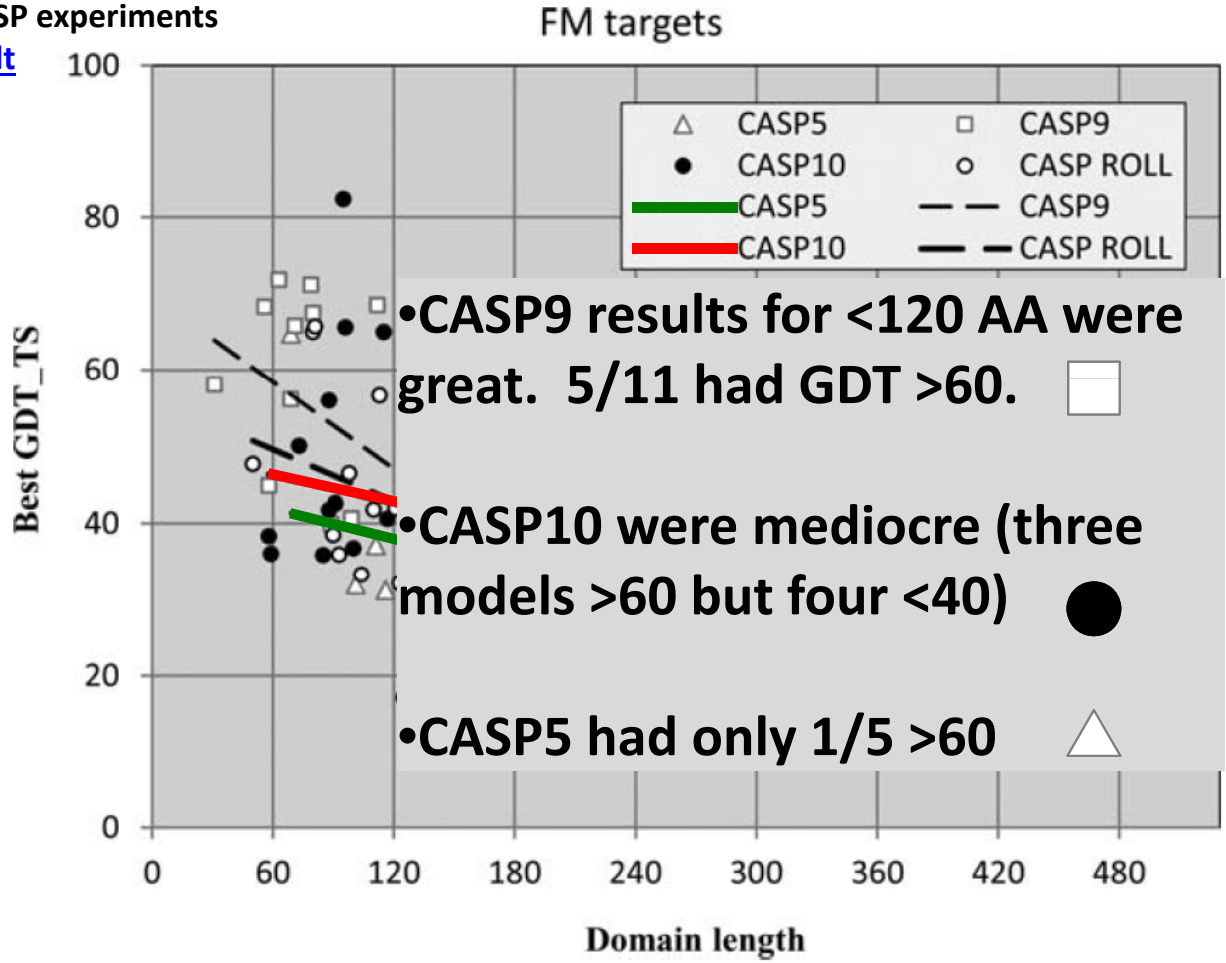
Free Modeling in Flux

CASP10 results compared to those of previous CASP experiments

Andriy Kryshtafovych, Krzysztof Fidelis, John Moult

DOI: 10.1002/prot.24448

GDT_TS
Perfect model: 90-100
Random model: 20-30



Source: Kryshtafovych, Andriy, Krzysztof Fidelis, et al. "CASP10 Results Compared to those of Previous CASP Experiments." *Proteins: Structure, Function, and Bioinformatics* 82, no. S2 (2014): 164-74.

Free Modeling in Flux

CASP10 results compared to those of previous CASP experiments
Andriy Kryshtafovych, Krzysztof Fidelis, John Moult
DOI: 10.1002/prot.24448

"Current FM [free modeling] methods perform best on single domain regular structures... The apparent lack of progress in CASP10 and ROLL compared with CASP5 probably again reflects the more difficult nature of CASP10 targets.

First, many targets which in CASP5 would have been in this category now have templates …

CASP10 FM targets exhibit more irregularity, and more of a tendency to be domains of larger proteins that are hard to identify from sequence and that may be dependent on the rest of the structure for their conformation.

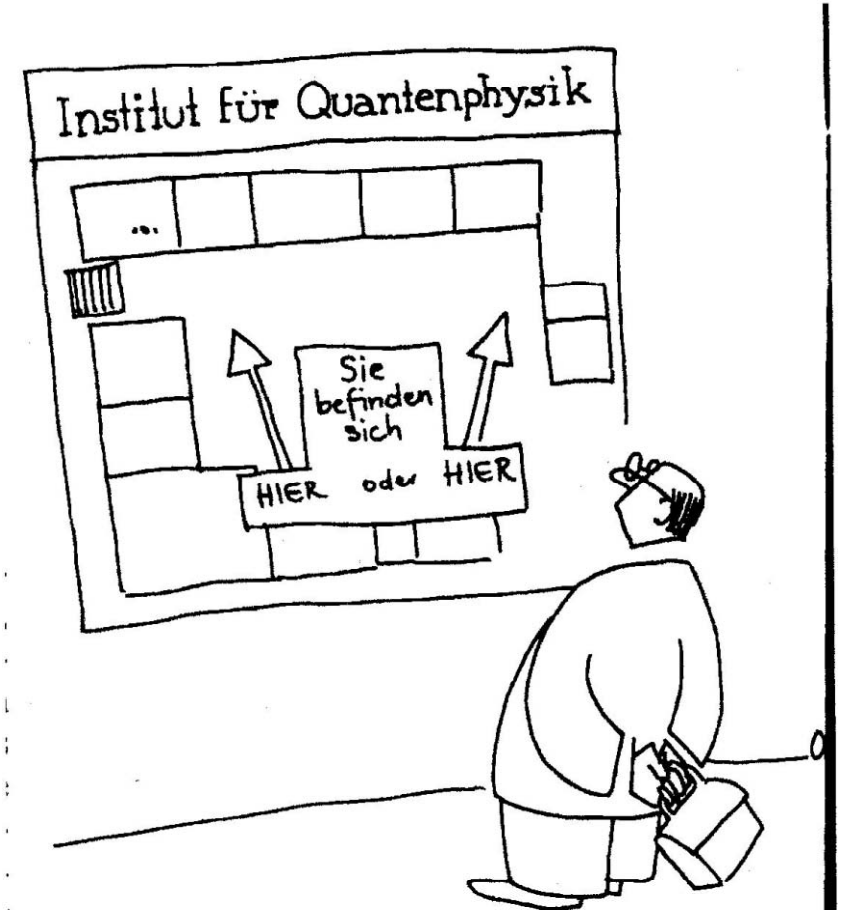
Statisticians vs. Physicists

"Data don't make any sense,
we will have to resort to statistics."

Rosetta

- Leverage everything we know about existing structures of proteins and peptides to build starting models
- Refine using a knowledge-based potential

Statisticians vs. Physicists

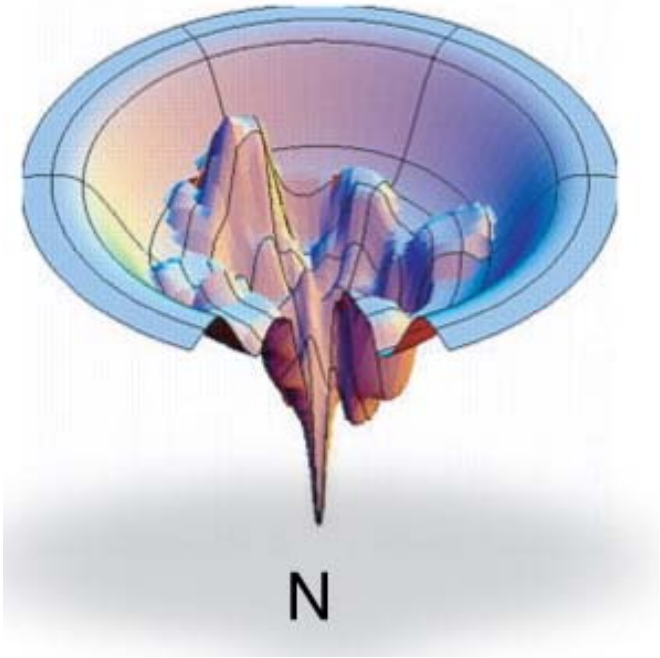


DE Shaw

- DON'T CHEAT!
- Only use physical forces.
- Fold proteins by simulating the in vitro process

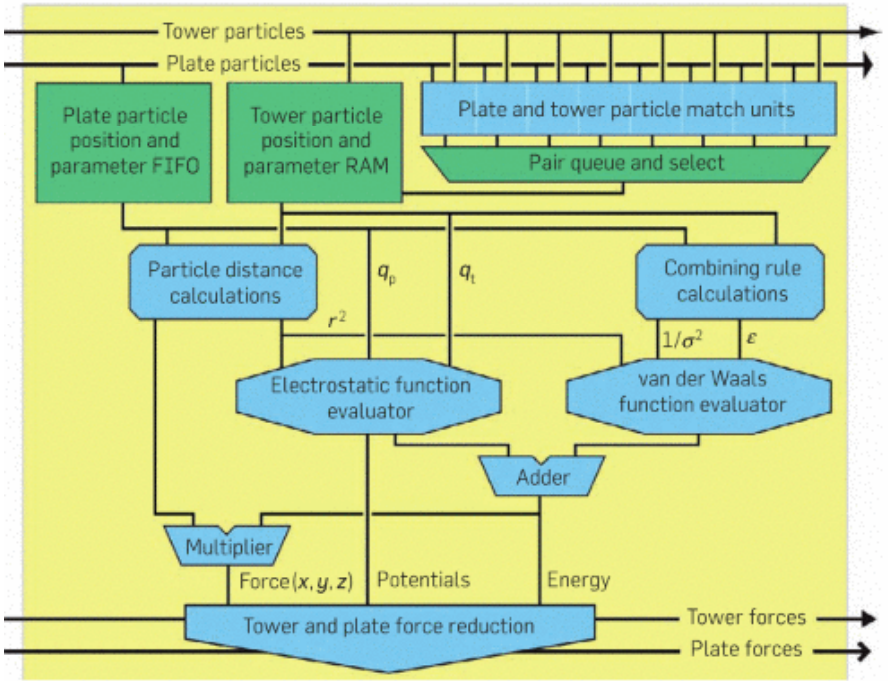
DE Shaw

- Lindorff-Larsen et al. (2011) Science
- Simulate protein folding.
- Why had no one else succeeded at this?

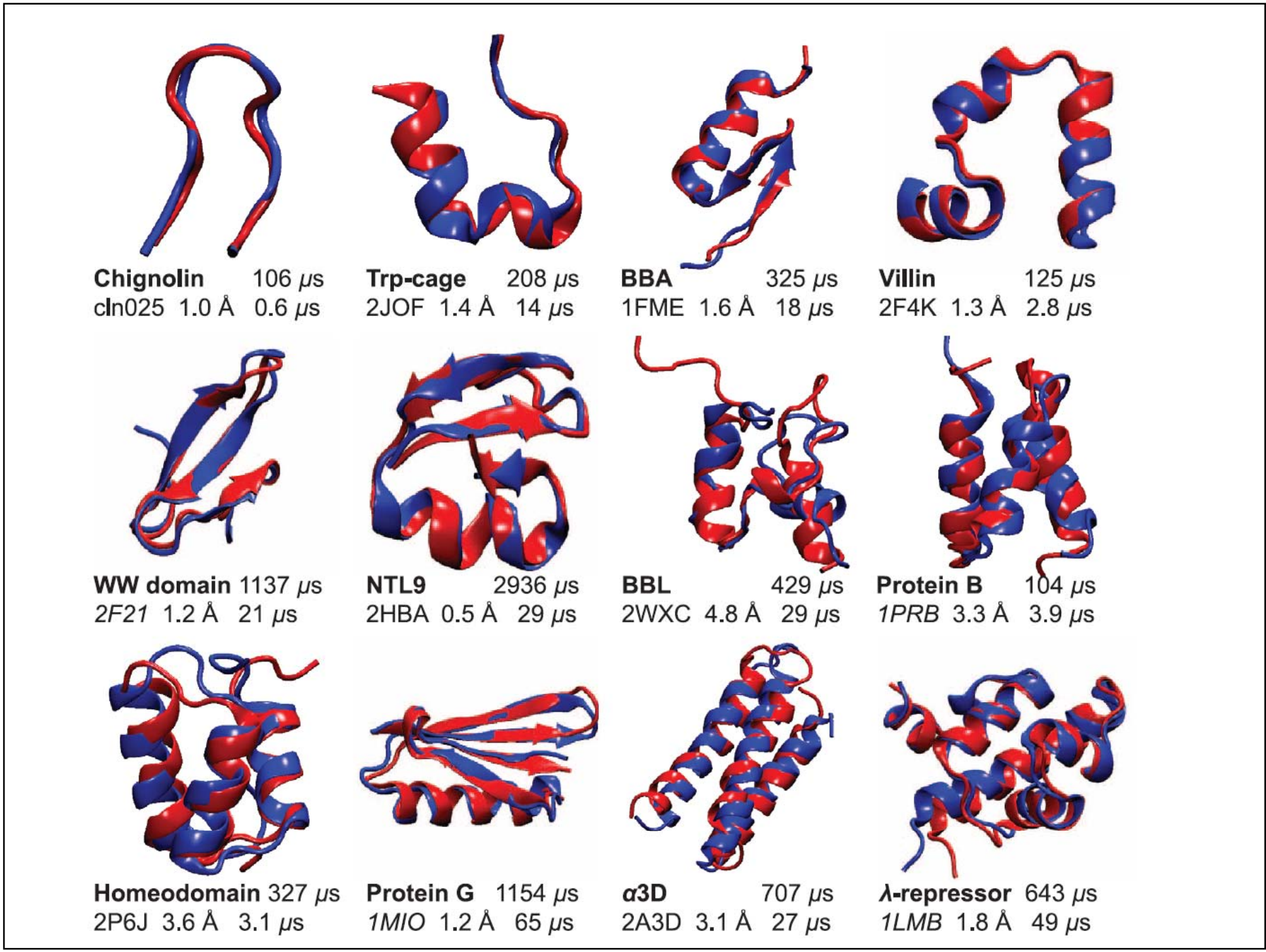


DE Shaw

- Lindorff-Larsen et al. (2011) Science
- Simulate protein folding.
- Built a specialized supercomputer
  - Hundreds of application specific integrated circuits



Source: Shaw, David E., Martin M. Deneroff, et al. "Anton, A Special-purpose Machine for Molecular Dynamics Simulation." *Communications of the ACM* 51, no. 7 (2008): 91-7.



Source: Lindorff-Larsen, Kresten, Stefano Piana, et al. "How Fast-folding Proteins Fold." *Science* 334, no. 6055 (2011): 517-20.

FoldIT Game

The New York Times

In a Video Game, Tackling the Complexities of Protein Folding

By JOHN MARKOFF
Published: August 9, 2010

Gamers 1, computer 0.

Enlarge This Image

PUZZLE University of Washington scientists developed Foldit, a free online game that drew thousands of players. "It's like trying to solve a million-sided Rubik's Cube while it also spins at 10,000 r.p.m.," a player wrote in a Web forum.

In a match that pitted video game players against the best known computer program designed for the task, the gamers outperformed the software in figuring out how 10 proteins fold into their three-dimensional configurations.

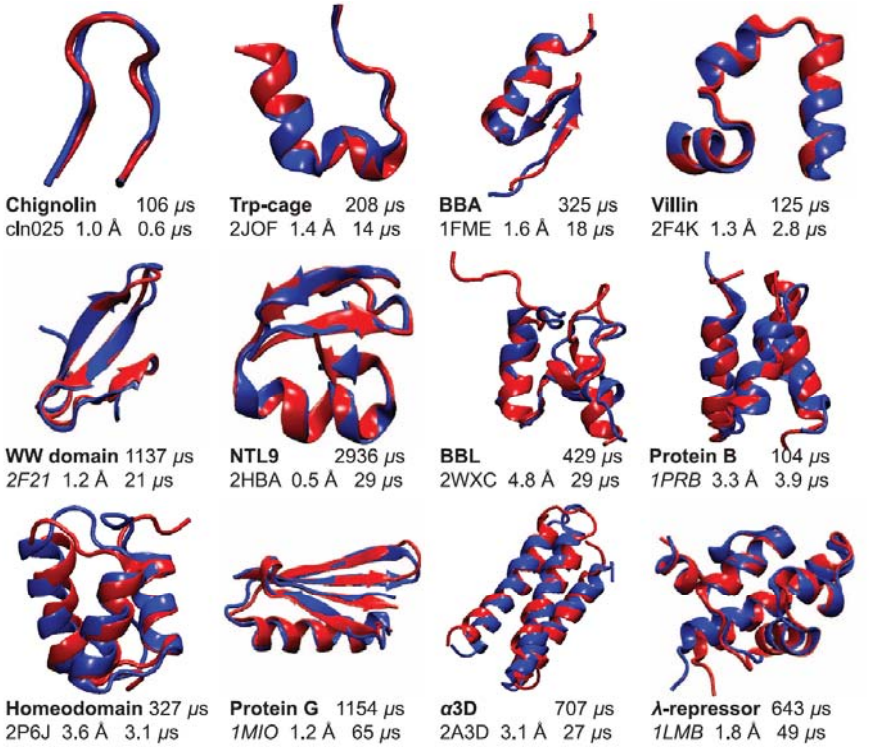
Proteins are essentially biological nanomachines that carry out myriad functions in the body, and biologists have long sought to understand how the long chains of

RECOMMEND
TWITTER
LINKEDIN
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Source: Markoff, John. "In a Video Game, Tackling the Complexities of Protein Folding." *The New York Times*. August 4, 2010.

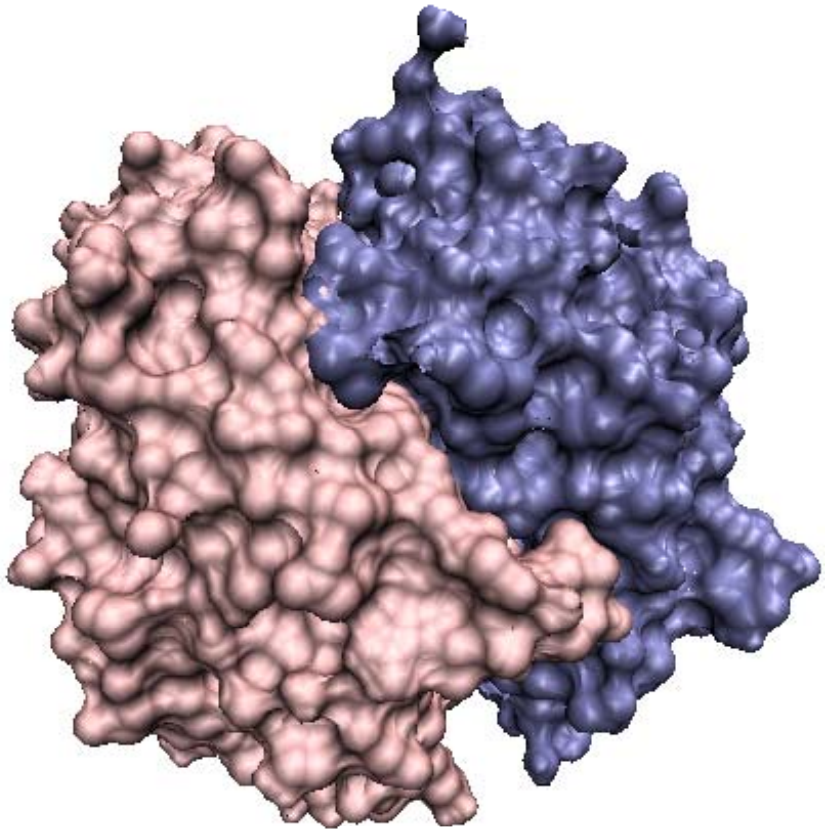
Predictions

So far: protein structure



Source: Lindorff-Larsen, Kresten, Stefano Piana, et al. "How Fast-folding Proteins Fold." *Science* 334, no. 6055 (2011): 517-20.

Next: protein interactions



Prediction Challenges

- Predict effect of point mutations
- Predict structure of complexes
- Predict all interacting proteins

Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions

•DOI: 10.1002/prot.24356

"Simple" challenge: Starting with known structure of a complex: predict how much a mutation changes binding affinity.

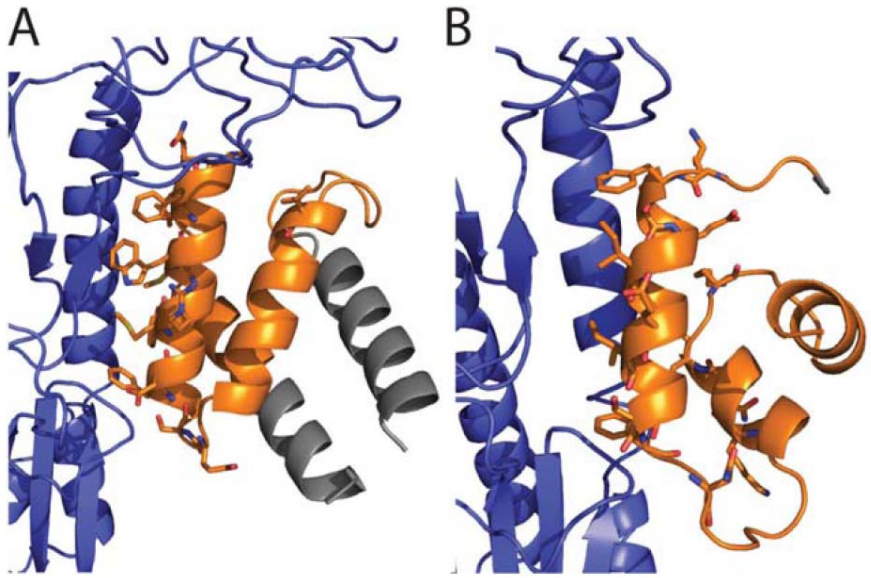


Figure 1

The structures of (**A**) HB36 (**B**) HB80 in complex with HA (blue) which were provided to participants. Residues probed in the deep sequencing enrichment experiment are in orange; the remainder are in grey. Residues at the interface are represented as sticks.

Source: Moretti, Rocco, Sarel J. Fleishman, et al. "Community-wide Evaluation of Methods for Predicting the Effect of Mutations on Protein–protein Interactions." *Proteins: Structure, Function, and Bioinformatics* 81, no. 11 (2013): 1980-7.

Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions

•DOI: 10.1002/prot.24356

•All possible single-point mutations at each of 53 and 45 positions for two proteins.

•Expressed on yeast

•High-throughput assay based on sequencing used to estimate changes in binding affinity

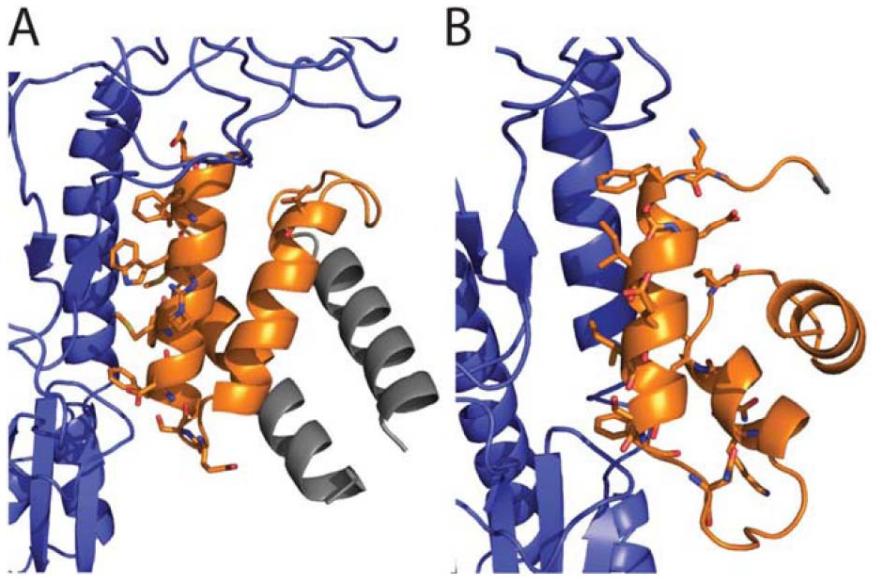


Figure 1

The structures of (**A**) HB36 (**B**) HB80 in complex with HA (blue) which were provided to participants. Residues probed in the deep sequencing enrichment experiment are in orange; the remainder are in grey. Residues at the interface are represented as sticks.

Source: Moretti, Rocco, Sarel J. Fleishman, et al. "Community-wide Evaluation of Methods for Predicting the Effect of Mutations on Protein–protein Interactions." *Proteins: Structure, Function, and Bioinformatics* 81, no. 11 (2013): 1980-7.

Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions

•DOI: 10.1002/prot.24356

How could we make quantitative predictions of binding energy for mutants?

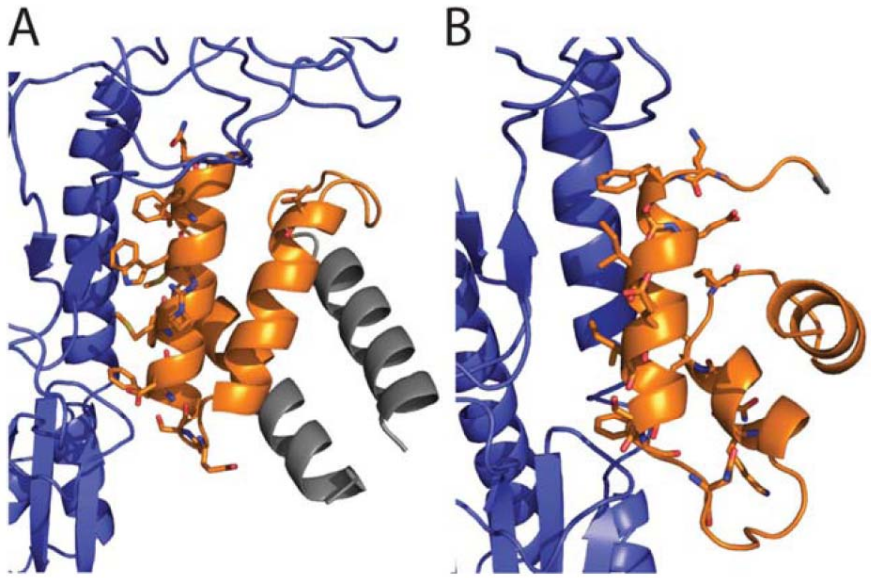
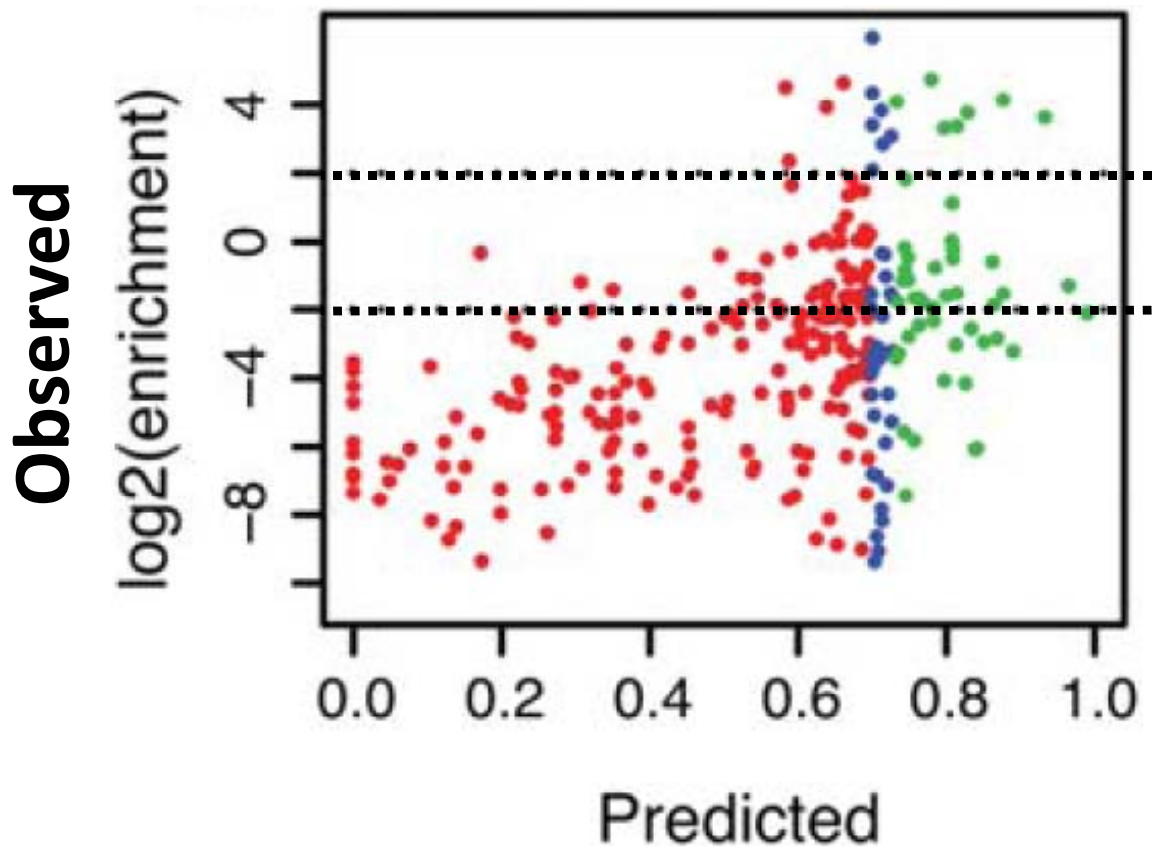


Figure 1

The structures of (**A**) HB36 (**B**) HB80 in complex with HA (blue) which were provided to participants. Residues probed in the deep sequencing enrichment experiment are in orange; the remainder are in grey. Residues at the interface are represented as sticks.



Color based on predictions

improved

neutral

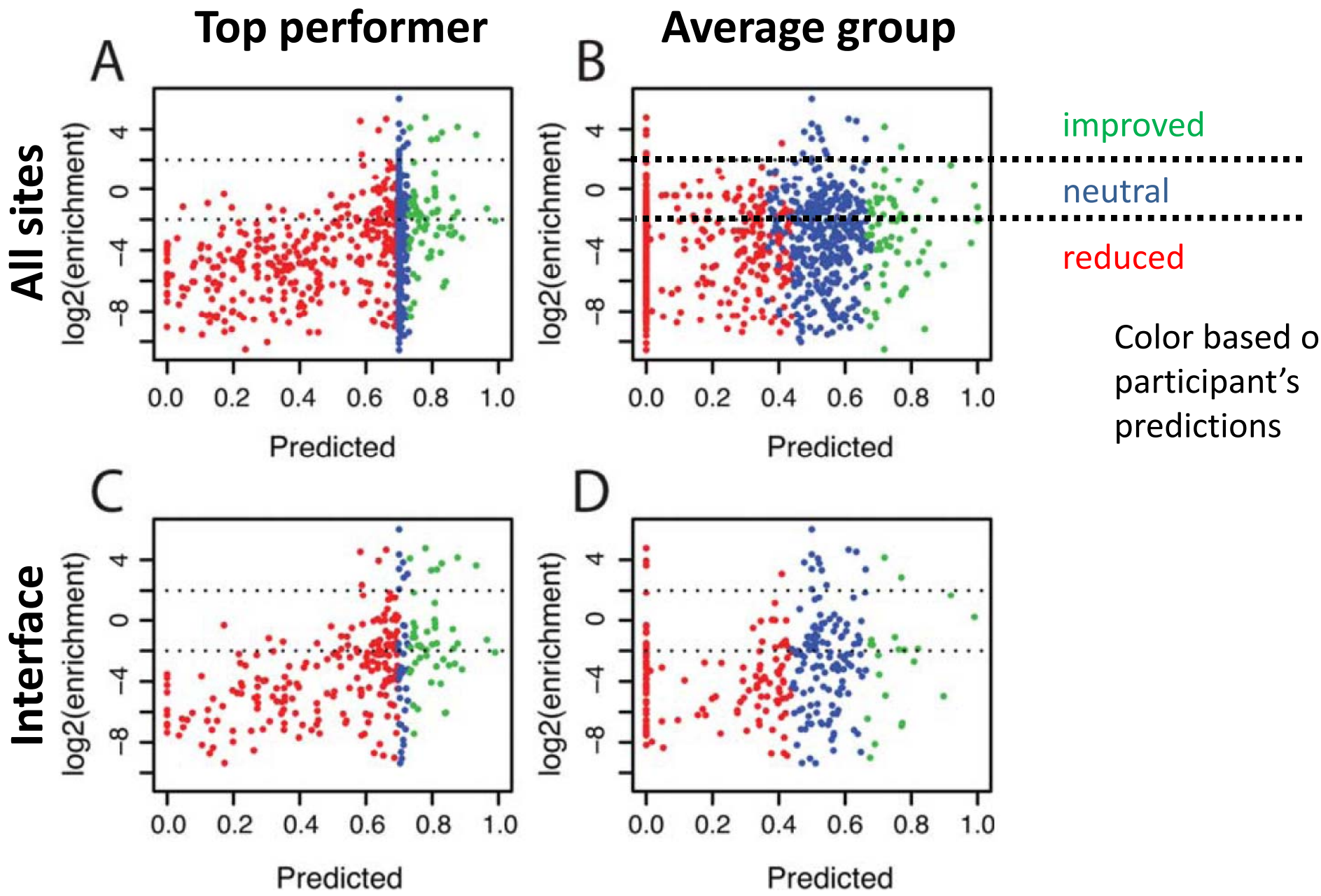
reduced

Note:
Better at predicting
deleterious mutations

Source: Moretti, Rocco, Sarel J. Fleishman, et al. "Community-wide Evaluation of Methods for Predicting the Effect of Mutations on Protein–protein Interactions." *Proteins: Structure, Function, and Bioinformatics* 81, no. 11 (2013): 1980-7.

This is one of the top performers analyzing residues at the interface!

•DOI: 10.1002/prot.24356



Source: Moretti, Rocco, Sarel J. Fleishman, et al. "Community-wide Evaluation of Methods for Predicting the Effect of Mutations on Protein–protein Interactions." *Proteins: Structure, Function, and Bioinformatics* 81, no. 11 (2013): 1980-7.

•DOI: 10.1002/prot.24356

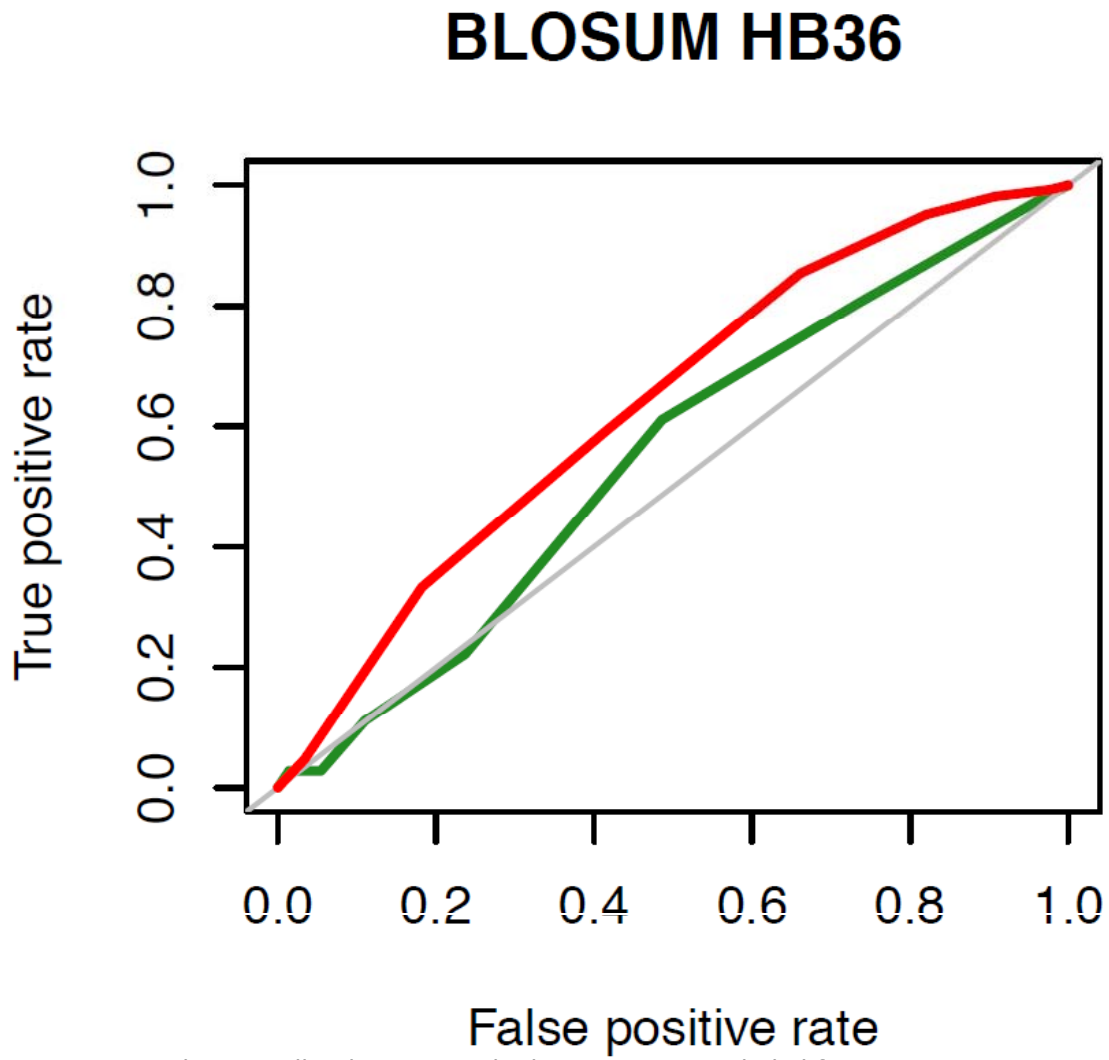
What's a good "baseline" for modeling?

- Does structure/energy help?

What's a good "baseline" for modeling?

- Does structure/energy help?
- Naïve model:
  - Give each mutant a score equal to the BLOSUM matrix value (-4 to 11)
  - As we vary the cutoff, how many mutations do we predict correctly?

Area under curve for predictions (varying cutoff in ranking)

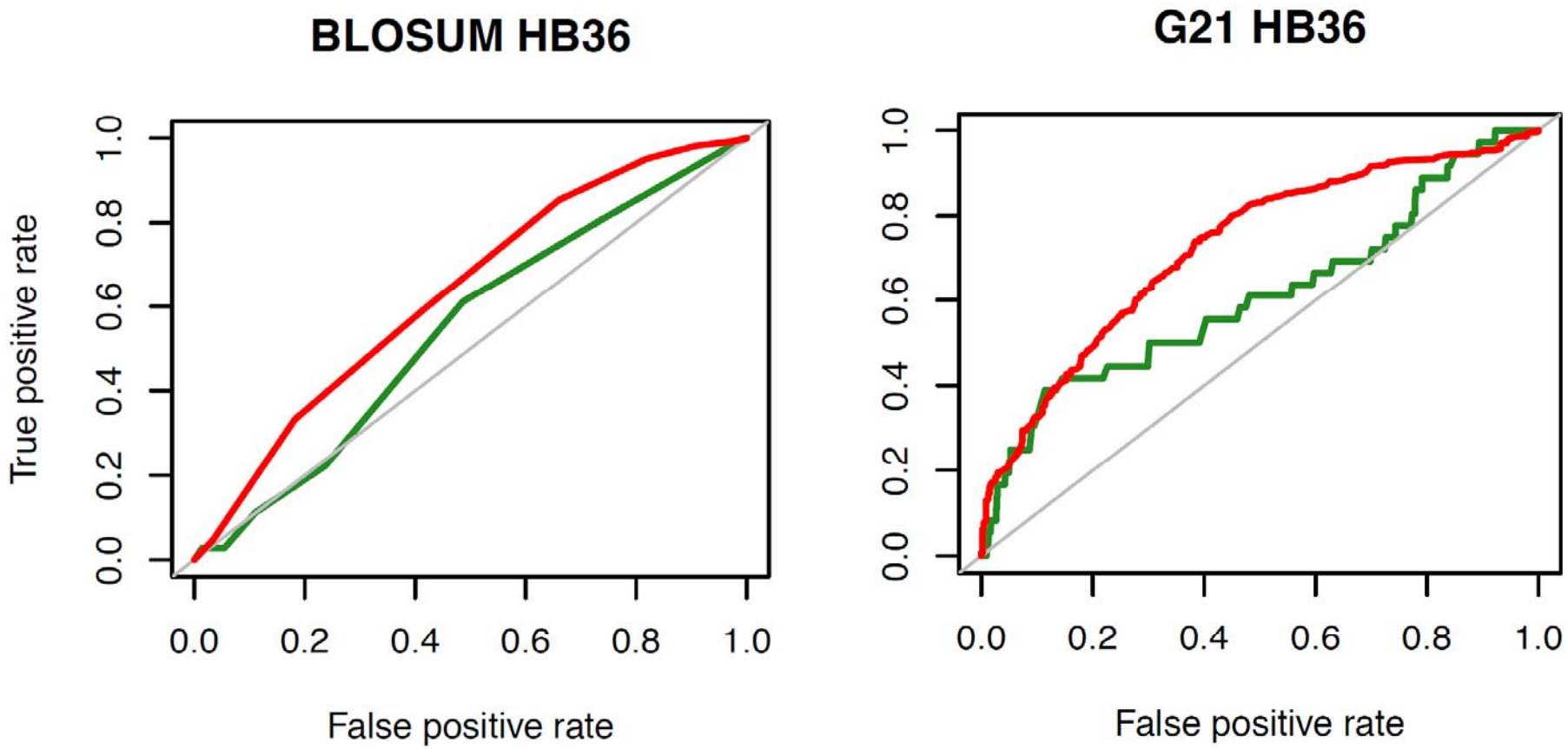


Predicted to be deleterious

Predicted to be beneficial

•DOI: 10.1002/prot.24356

Comparing one of the best to BLOSUM

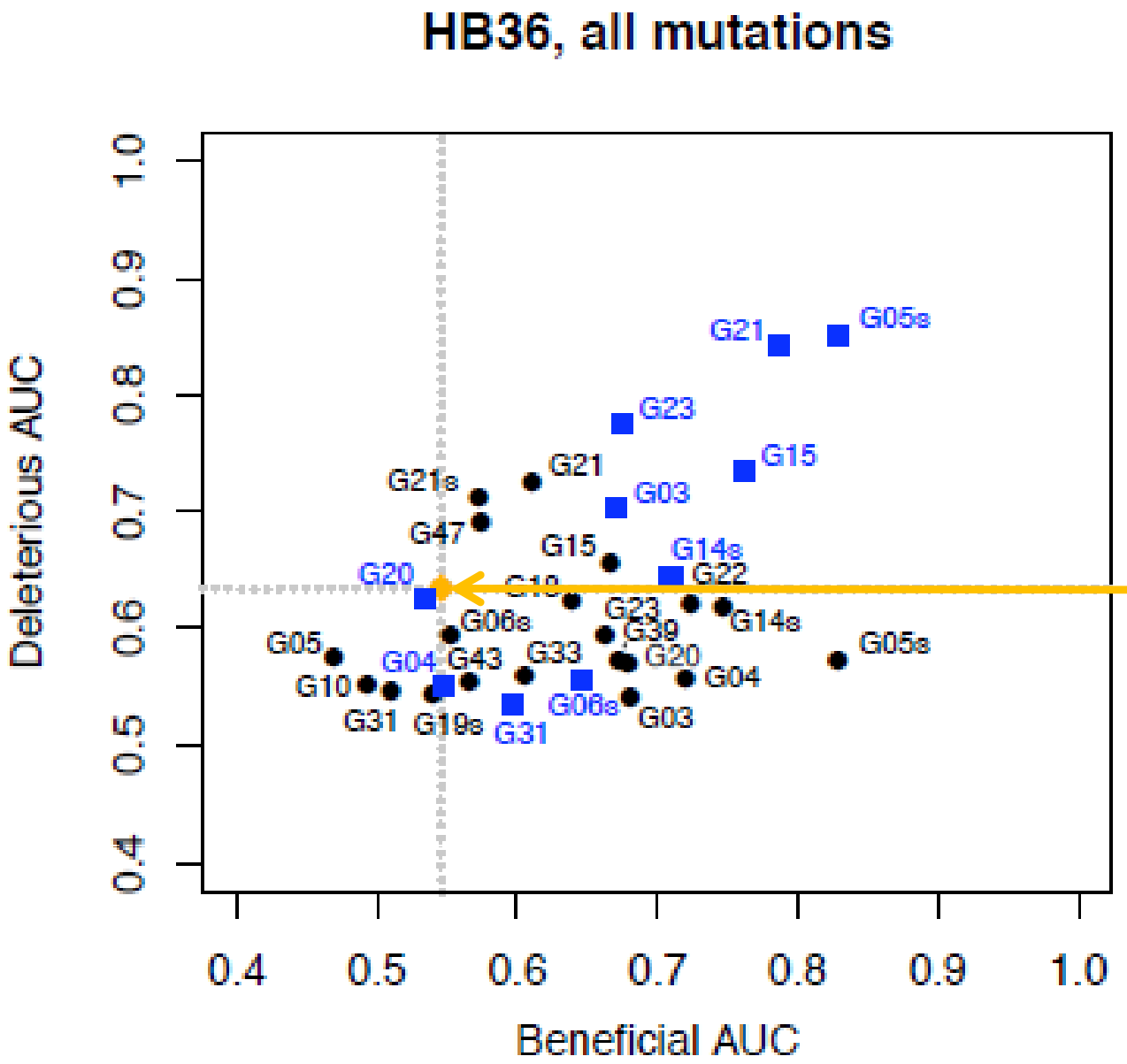


Predicted to be deleterious

Predicted to be beneficial

•DOI: 10.1002/prot.24356

Area under curve for predictions (varying cutoff in ranking)



First Round

Second Round. (Given data for nine random mutations at each position)

BLOSUM

•DOI: 10.1002/prot.24356

Summary

- Best groups are only three times better than expected from a random assignment.
- Predicting the effect of mutations on polar starting positions appears to be a particular challenge.

Summary

- Best approaches require explicit consideration of the effects of mutations on stability

$$A + B \rightleftarrows AB$$

$$A + B^{*} \rightleftarrows AB^{*}$$

Summary

- Best approaches require explicit consideration of the effects of mutations on stability

$$\text{unfolded} \rightleftarrows A + B \rightleftarrows AB$$

$$\text{unfolded} \rightleftarrows A + B^{*} \rightleftarrows AB^{*}$$

For more details see

http://ocw.mit.edu/courses/biological-engineering/20-320-analysis-of-biomolecular-and-cellular-systems-fall-2012/modeling-and-manipulating-biomolecular-interactions/MIT20_320F12_Tpc_3_Mol_Des.pdf

Summary

- Best approaches require explicit consideration of the effects of mutations on stability
- The best performing groups also modeled packing, electrostatics and solvation.
- The best methods used :
  - machine learning (G21, Fernandez-Recio, and G05s, Bates)
  - atom-level energy functions (G15, Weng)
  - coarse-grained models (G21s, Dehouck)

G21

- Database of 930 (ΔΔG,mutation) pairs
- Predict structure with FoldX (empirical force field)
- Describe each mutant with 85 features using measures from FoldX, PyRosetta, FireDoc, PyDoc, SIPPER, CHARMM, NIP/NSC and others.

G21

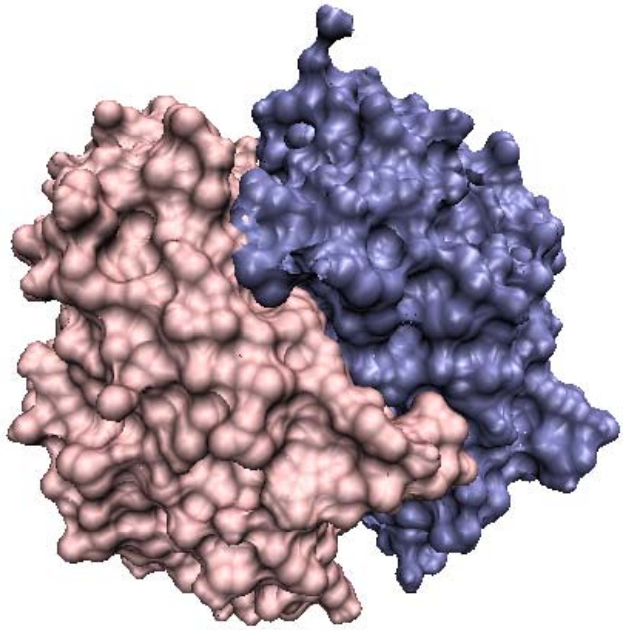
- Train learners: random forest, neural networks, probabilistic classifiers, etc.
- Evaluate with cross-validation
- Use combined results from five classifiers:
  - Random forest
  - Decision table
  - Bayesian net
  - Logistic regression
  - Alternating decision tree

Prediction Challenges

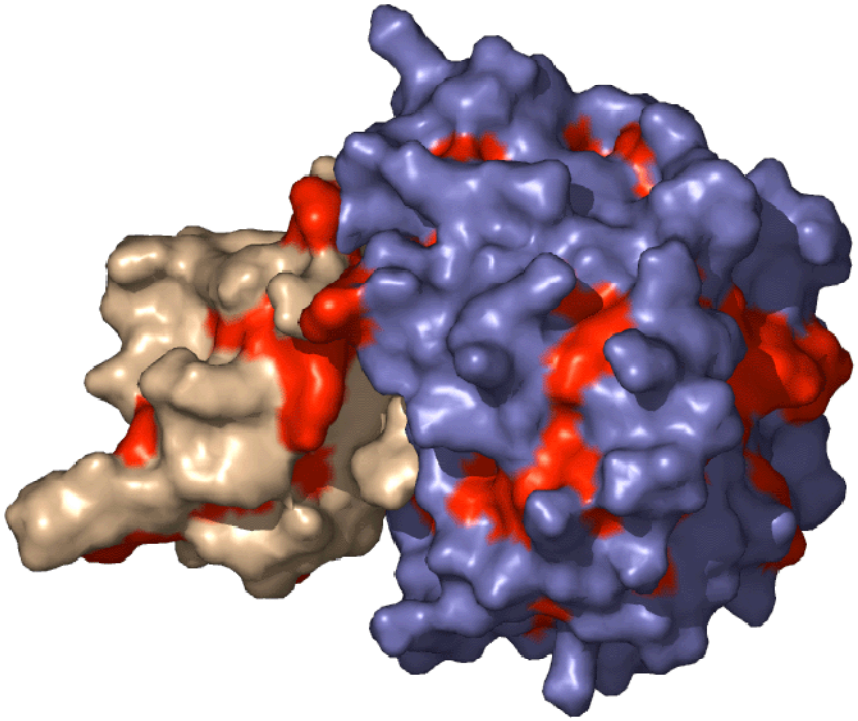
- Predict effect of point mutations
- Predict structure of complexes
- Predict all interacting proteins

Predicting Structures of Complexes

- Can we use structural data to predict complexes?
- This might be easier than quantitative predictions for site mutants.
- But it requires us to solve a docking problem



Docking



Courtesy of Nurcan Tuncbag. Used with permission.

Which surface(s) of protein A interactions with which surface of protein B?

N. Tuncbag

Time is an issue

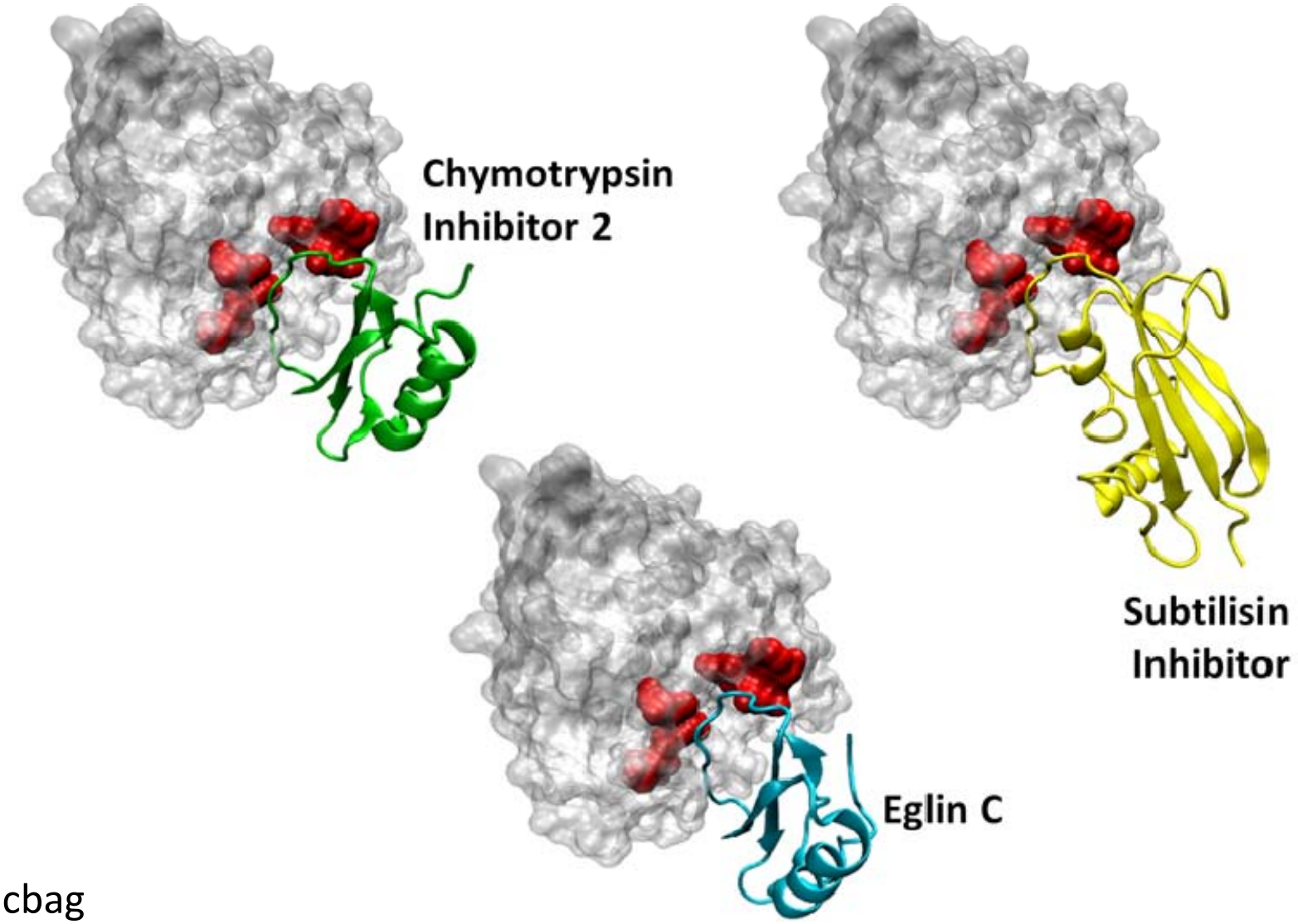
- Imagine we wanted to predict which proteins interact with our favorite molecule.
  - For each potential partner
    - Evaluate all possible relative positions and orientations
      - allow for structural rearrangements
        - » measure energy of interaction
- This approach would be extremely slow!
- It's also prone to false positives.
  - Why?

Reducing the search space

- Use prior knowledge of interfaces to focus analysis on particular residues
- Find ways to choose potential partners
  - What role should structural homology play?

Subtilisin and its inhibitors

Although global folds of Subtilisin's partners are very different, binding regions are structurally very conserved.



N. Tuncbag

Courtesy of Nurcan Tuncbag. Used with permission.

Hotspots

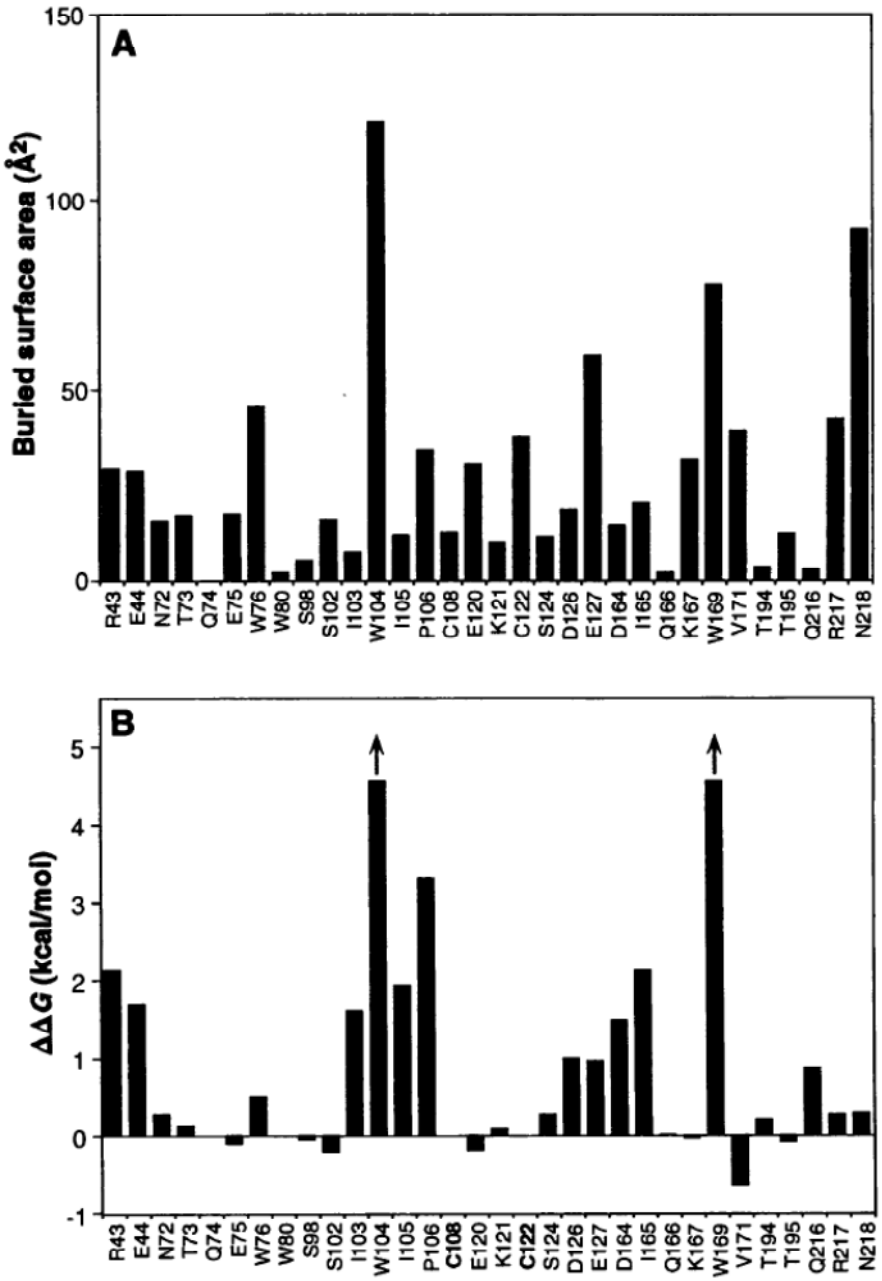
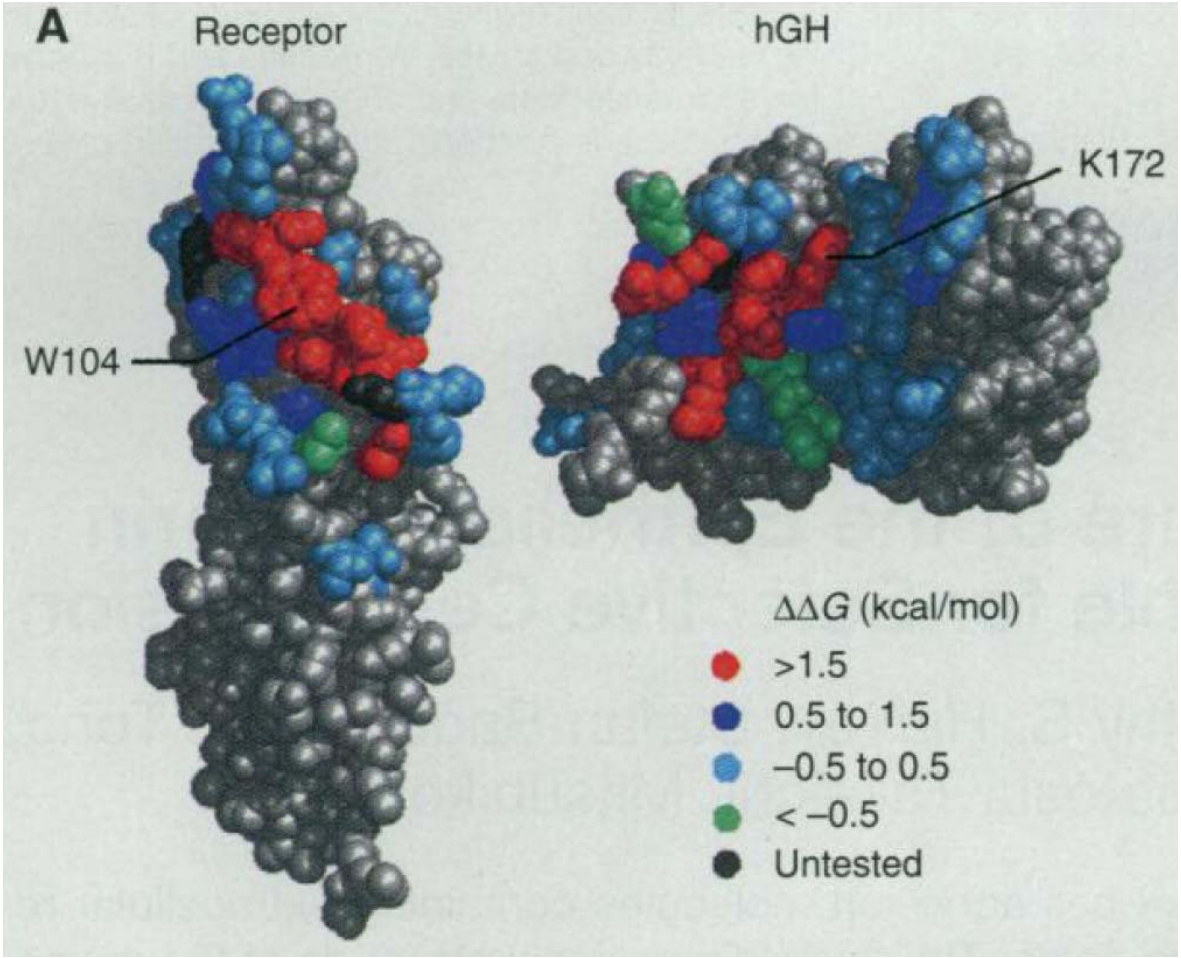


Fig. 1. Contribution of only a subset of contact residues to net binding energy. (**A**) Loss of solvent-accessible area (7) of the side chain portion of each residue in the hGHbp on forming a complex with hGH. (**B**) Difference in binding free energy between alanine-substituted and wild-type hGHbp $(\Delta\Delta G)_{mut\text{-}wt}$ at contact residues (5). Negative values indicate that affinity increased when the side chain was substituted by alanine.

Figure from Clackson & Wells (1995).

Source: Clackson, Tim, and James A. Wells. "A Hot Spot of Binding Energy in a Hormone-Receptor Interface." *Science* 267, no. 5196 (1995): 383-6.

Hotspots



Source: Clackson, Tim, and James A. Wells. "A Hot Spot of Binding Energy in a Hormone-Receptor Interface." *Science* 267, no. 5196 (1995): 383-6.

Figure from Clackson & Wells (1995).

- Fewer than 10% of the residues at an interface contribute more than 2 kcal/mol to binding.
- Hot spots
  - rich in Trp, Arg and Tyr
  - occur on pockets on the two proteins that have complementary shapes and distributions of charged and hydrophobic residues.
  - can include buried charge residues far from solvent
  - O-ring structure excludes solvent from interface

http://onlinelibrary.wiley.com/doi/10.1002/prot.21396/full

Next Lecture

Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement.

Tuncbag N, Keskin O, Nussinov R, Gursoy A.

http://www.ncbi.nlm.nih.gov/pubmed/22275112

Structure-based prediction of protein–protein interactions on a genome-wide scale

Zhang, et al.

http://www.nature.com/nature/journal/v490/n7421/full/nature11503.html

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014