



Computational Complexity in Biology

Teresa Head-Gordon
Physical Biosciences and Life Sciences Divisions
Lawrence Berkeley National Laboratory

April 5, 2000

Why Computational Biology?



(1) Community effort to define problems with genuine computational complexity

Genome analysis, gene modeling, sequence-based annotation

Low resolution fold prediction: Single Molecule

High resolution structure prediction and protein folding: Single Molecule

Molecular recognition or Docking: Multi-molecule complexes

Cellular Decision modeling

(2) Putting it all together

Deinococcus radiodurans

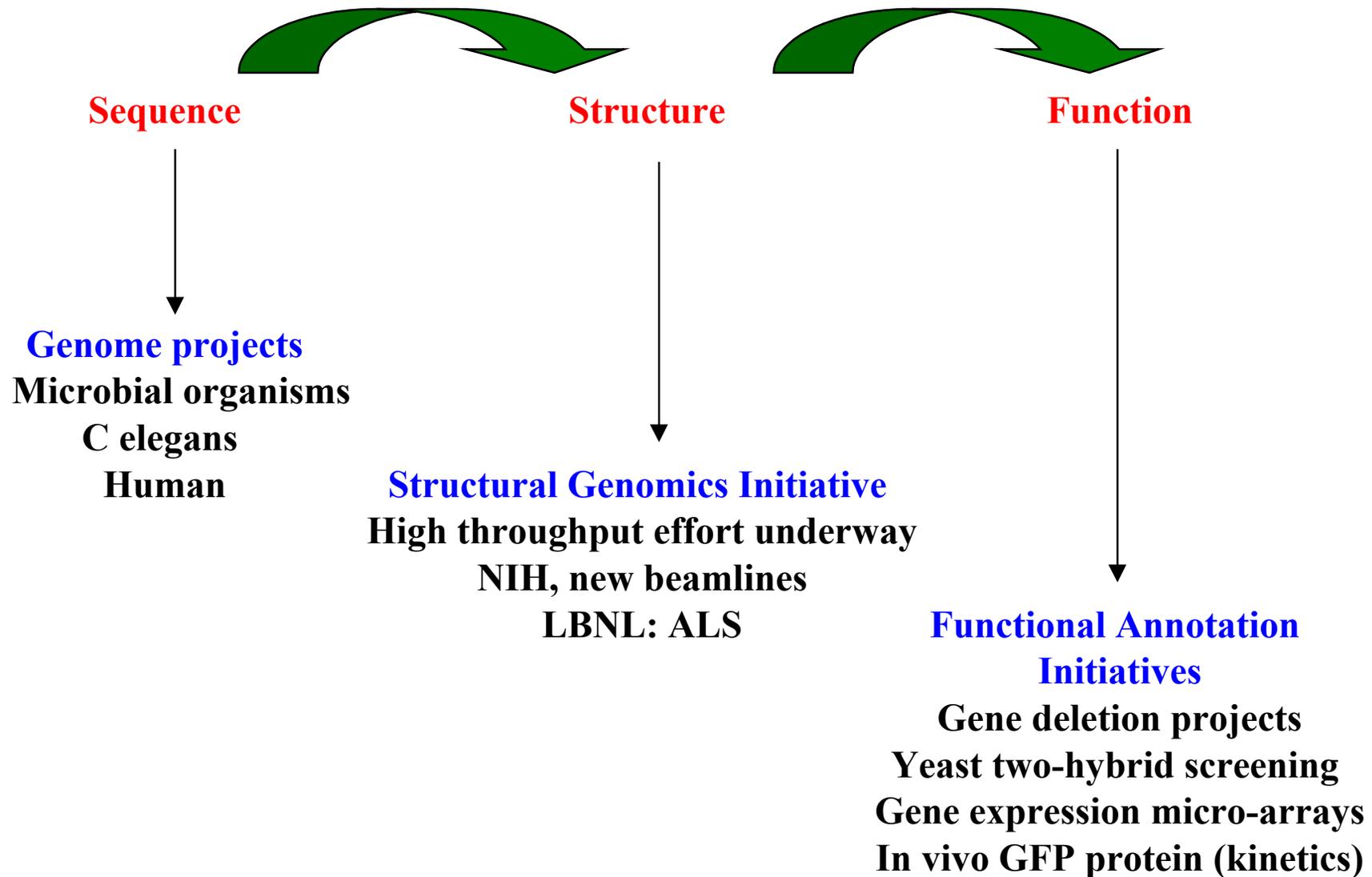
CASP competition for protein structure prediction

(3) Research examples from THG group

Global Optimization approach to predict protein structure

Simulation/experiment to understand hydration in protein folding

Revolutionary Experimental Efforts in Biology



Computational Biology White Paper



<http://cbcg.lbl.gov/ssi-csb>

Technical document to define areas of computational biology problems of scale

Organization:

Introduction to biological complexity, needs for advanced computing (1)

Scientific areas (2-6)

Computing hardware, software, CSET issues (7)

Appendices

For each scientific chapter:

illustrate with state of the art application (current hpc platform)

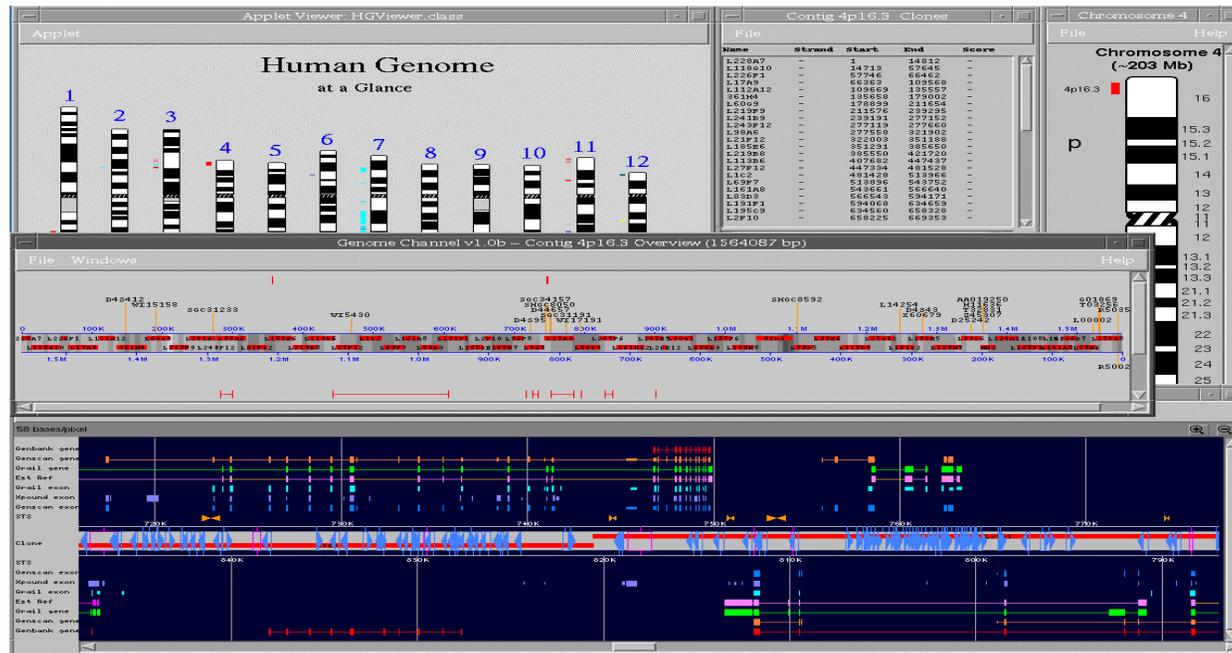
define algorithmic kernals

deficiencies of methodologies

define what can be accomplished with 100 teraflop computing

- **Community document**
- **More organized CB community in government labs, universities**
 - **Support for CB by the broader biological community**

High-Throughput Genome Sequence Assembly, Modeling, and Annotation



The Genome Channel Browser to access and visualize current data flow, analysis and modeling. (Manfred Zorn, NERSC)



Genome sequencing and annotation ———> **Bioinformatics**

100,000 human genes; genes from other organism

Structure/functional annotation at the sequence level

Computation to determine regions of a genome that might yield new folds

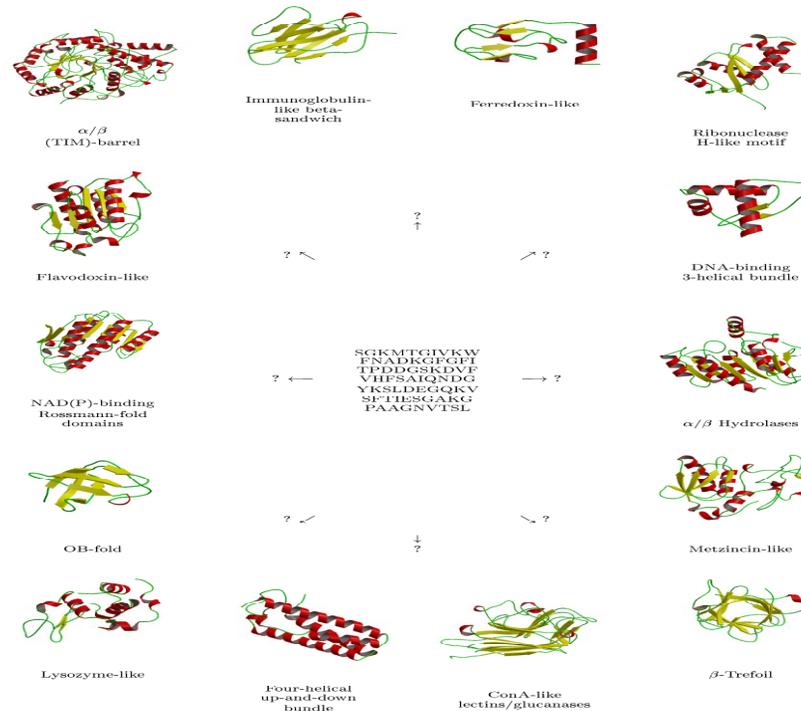
Experimental Structural Genomics Initiative

Functional annotation at the structure level by experiment

Characterize the link between protein sequence and fold topology



Sequence Assignments to Protein Fold Topology (David Eisenberg, UCLA)



Experimental Structural Genomics Initiative

Define basis set of folds: $\sim 10^3$ structures to be determined

Predict Fold Topology from Computation ($\sim 10^5$ folds)

Functional annotation at the structural level by computation

Low Resolution Fold Topologies to High Resolution Structure

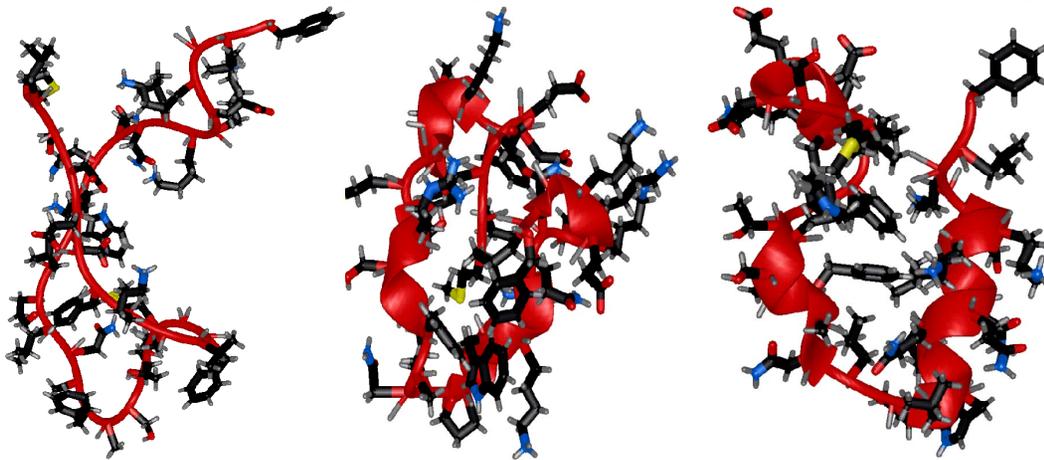


Low Resolution Structures from Predicted Fold Topology

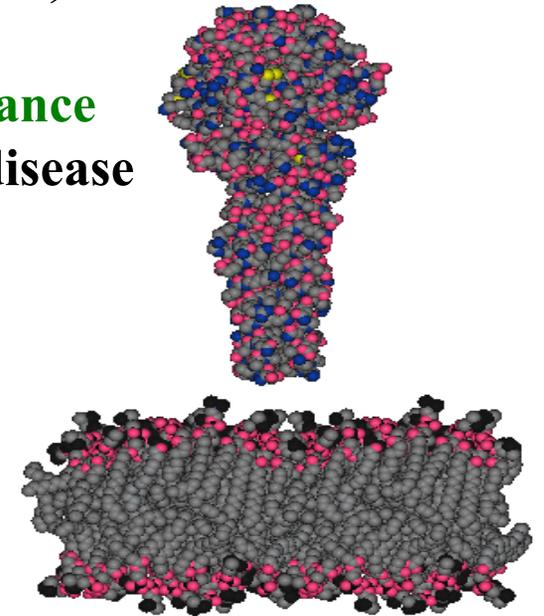
Fold class gives some idea of biological function, but....

Higher Resolution Structures with Biochemical Relevance

Drug design, bioremediation, new pathogen disease

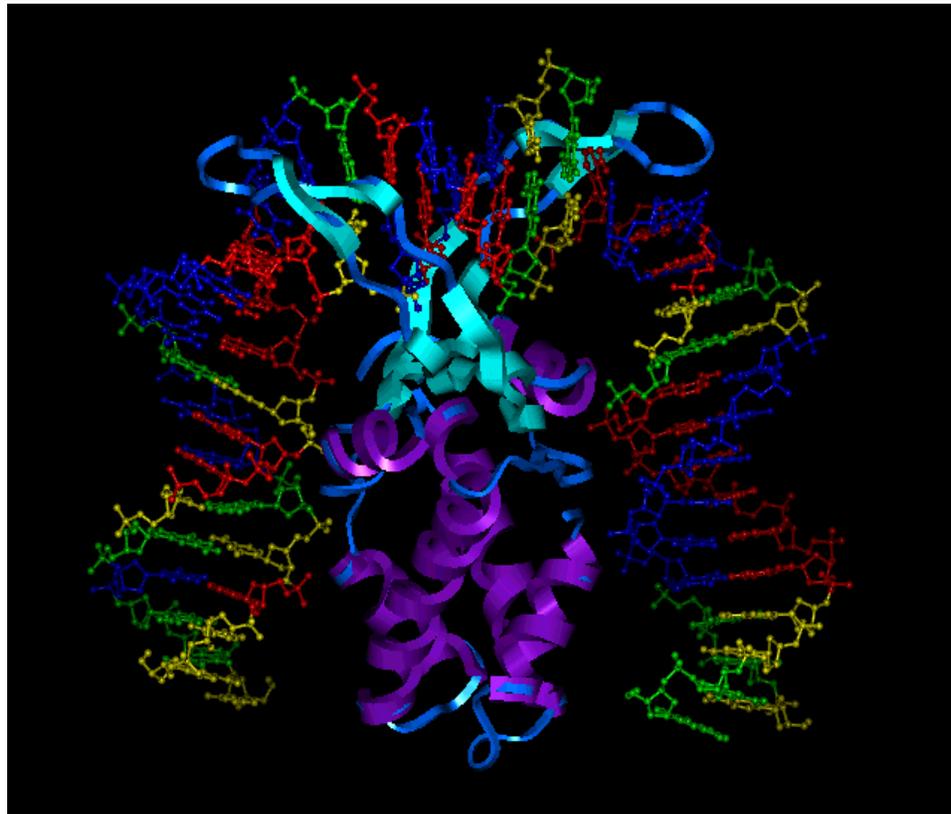


*One microsecond simulation of a fragment of small protein
Villin headpiece
Duan & Kollman, Science 1998*



*Influenza virus poised above
model of a lipid membrane will
involve a 100,000 atom MD
simulation over long timescales
to understand this step in the
mechanism of viral infection.
(Tobias, UCI)*

Simulating Molecular Recognition/Docking



Changes in the structure of DNA that can be induced by proteins.

Through such mechanisms proteins regulate genes, repair DNA, and carry out other cellular functions.

Improvements in Methodology and Algorithms of Higher Resolution Structure

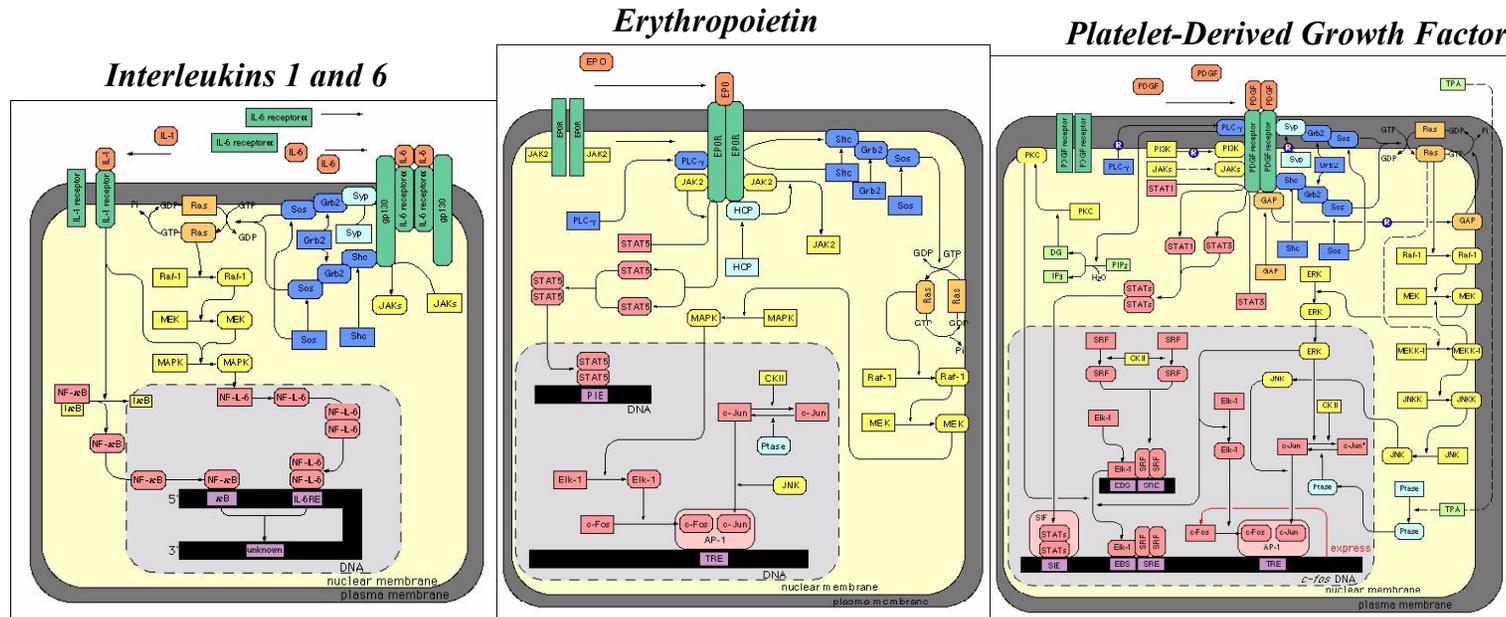
Break down size, time, lengthscale bottlenecks

Protein, DNA recognition, binding affinity, mechanism of drug binding

Simulating two-hybrid yeast experiments

Protein-protein and Protein-nucleic acid docking

Modeling the Cellular Program



Three mammalian signal transduction pathway that share common molecular elements (i.e. they cross-talk). From the Signaling Pathway Database (SPAD) (<http://www.grt.kyushu-u.ac.jp/spad/>)



Integrating Computational/Experimental Data at all levels

Sequence, structural functional annotation

Simulating biochemical/genetic networks to model cellular decisions

Modeling of network connectivity (sets of reactions: proteins, small molecules, DNA)

Functional analysis of that network (kinetics of the interactions)

Implicit Collaborations with Computer Science



Computer Hardware & Portability

Applications described running on various platforms

T3D, T3E, IBM SP's, ASCI Red, Blue

Information Technologies and Database Management

Integrating biological databases; CORBA and java

Data Warehousing

ultra-high-speed networks

Ensuring Scalability on Parallel Architectures

implicit algorithmic scaling

paradigm/software library support tools for effective parallelization

strategies: 100 teraflop

Meta Problem Solving Environments

geographically distributed software paradigm: “plug and play” paradigm

Visualization

Querying data which is “information dense”

Feedback from Biotech Industry Meeting



LBNL 2/25/99

Jim Cavalcoli, Ph.D.
Bioinform. Manager, PDLMG
Parke-Davis, Warner-Lambert

Patrick O'Hara
VP, Bioinformatics
ZymoGenetics, Inc
Seattle WA

Herve Recipon
Asst. Dir. bioinformatics
diaDexus (Incyte)

Pete Smietana, Ph.D.
Sr. Software Engineer,
Bioinformatics
Ciphergen

Peter Karp, Ph.D.
Scientific Fellow
Pangea Systems

Rick Bott
X-ray crystallographer
Genencor

Julie Rice
Computational Chemist
IBM-Almaden

Eric Martin
Sr. Scientist Small Molecule Discovery
Chiron

LBNL: Gilbert, Head-Gordon, Holbrook, Mian, Rokhsar, Simon, Spengler, Zorn

Biotech industry perspective on Computational Biology white paper

Is there strong objection to any of the content?

NO, very supportive

Are there other areas to be included, stronger emphasis placed?

Databases: integrating, querying, visualization

Deinococcus Radiodurans

Strange berry that withstands radiation



Bacteria isolated from tins of spoiled meat given “sterilizing” doses of γ radiation.
3x10⁶ base pairs, or ~3000 protein products
fully sequenced by TIGR under DOE/OBER sponsorship

Three components to DR's successful DNA repair strategy

specifics of the DNA repair mechanism

the fact that it is multi-genomic

coupling of repair, replication, export of damaged DNA from intracellular medium.

Construct molecular models of key components of the DNA repair system:

Damaged DNA

Multigenomic repair intermediates such as Holliday junctions

Proteins known are yet to be discovered to be involved in DNA repair

Protein-protein or protein-nucleic acids that couple repair, replication, transport.

Develop better fold recognition, comparative modeling, and ab initio prediction methods, and docking methods to describe macromolecular complexes.

Application of methodologies will be to fully and completely annotate the DR genome

Learn underlying components of highly-honed strategies for DNA repair in DR.

Significant portions of community white paper on high end computing needs

The Need for Advanced Computing for Computational Biology



Computational Complexity arises from inherent factors:

100,000 gene products just from human; genes from many other organisms

Experimental data is accumulating rapidly

N^2 , N^3 , N^4 , etc. interactions between gene products

Combinatorial libraries of potential drugs/ligands

New materials that elaborate on native gene products from many organisms

Algorithmic Issues to make it tractable

Objective Functions

Optimization

Treatment of Long-ranged Interactions

Overcoming Size and Time scale bottlenecks

Statistics

Acknowledgements for Community White Paper in Computational Biology



The First Step Beyond the Genome Project: High-Throughput Genome Assembly, Modeling, and Annotation

P. LaCascio, R. Mural, J. Snoddy, E. Uberbacher, ORNL
S. Mian, F. Olken, S. Spengler, M. Zorn: LBNL
David States, Washington University

From Genome Annotation to Protein Folds: Comparative Modeling and Fold Assignment

D. Eisenberg, UCLA
A. Lapedes, LANL
A. Sali, Rockefeller University
B. Honig, Columbia University

Low Resolution Folds to High Resolution Protein Structure and Dynamics

C. Brooks, Scripps Research Institute
P. Kollman & Y. Duan, UCSF
A. McCammon & V. Helms, UCSD
G. Martyna, Indiana University
D. Tobias, UCI
T. Head-Gordon, LBNL

Biotechnology Advances from Computational Structural Genomics: In Silico Drug Design and Mechanistic Enzymology

R. Abagyan, NYU, Skirball Institute
P. Bash, ANL
J. Blaney, Metaphorics, Inc.
F. Cohen, UCSF
M. Colvin, LLNL
I. Kuntz, UCSF

Linking Structural Genomics to Systems Modeling: Modeling the Cellular Program

A. Arkin & D. Wolf, LBNL
P. Karp, PangeaS. Subramaniam, U Illinois
Urbana

Implicit Collaborations Across the DOE Mission Sciences

M. Colvin & C. Musick, LLNL
T. Gaasterland, ANL (now Rockefeller)
S. Crivelli & T. Head-Gordon, LBNL
G. Martyna, Indiana University



Protein Fold Recognition, Structure Prediction, and Folding

Teresa Head-Gordon
Physical Biosciences and Life Sciences Divisions
Lawrence Berkeley National Laboratory

April 5, 2000

Computational Challenges in Biology

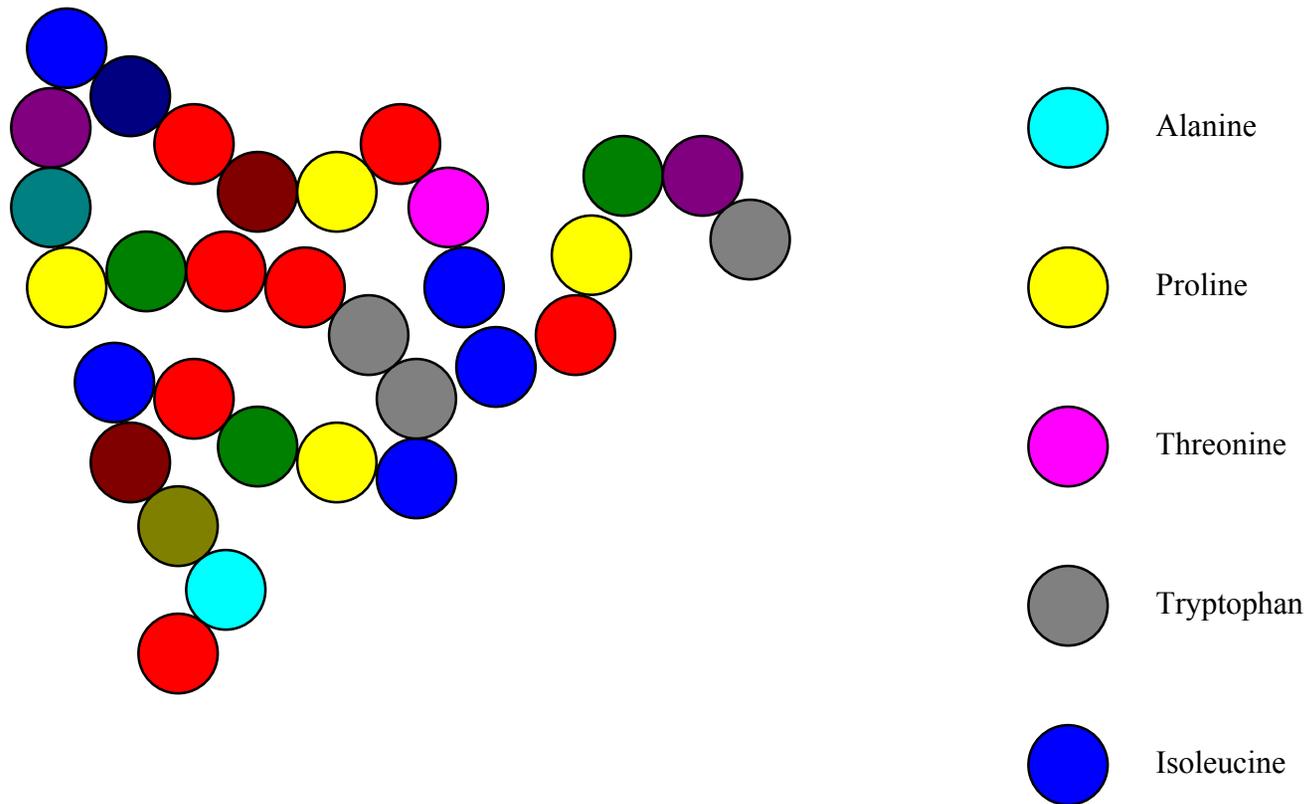


- (1) Drawing analogies with known protein structures
Sequence homology, Structural Homology
Inverse Folding, Threading
- (2) Ab initio prediction: the ability to extrapolate to unknown folds
multiple minima problem; robust objective function
- (3) Ab initio folding: the ability to follow kinetics, mechanism
robust objective function
severe time-scale problem
proper treatment of long-ranged interactions
- (4) Ab initio prediction: Global Optimization Approaches
Stochastic Perturbation and Soft Constraints
- (5) Simplified Models that Capture the Essence of Real Proteins
Off-Lattice Model that Connect to Experiments: Whole Genomes?
- (6) Simulating experimental observables
Hydration forces in folding

What is a protein?

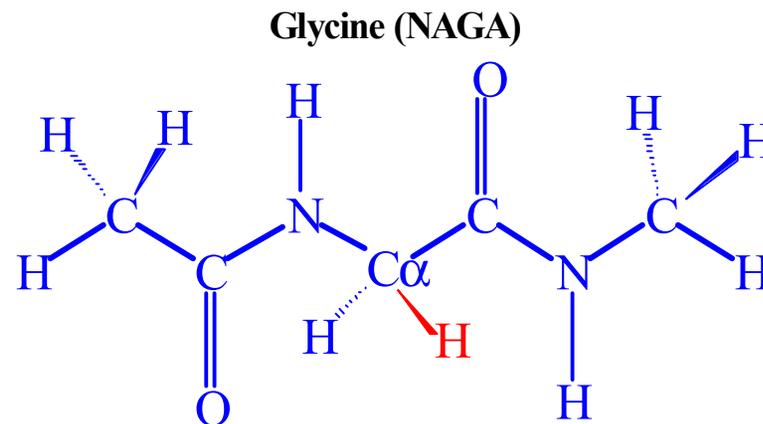
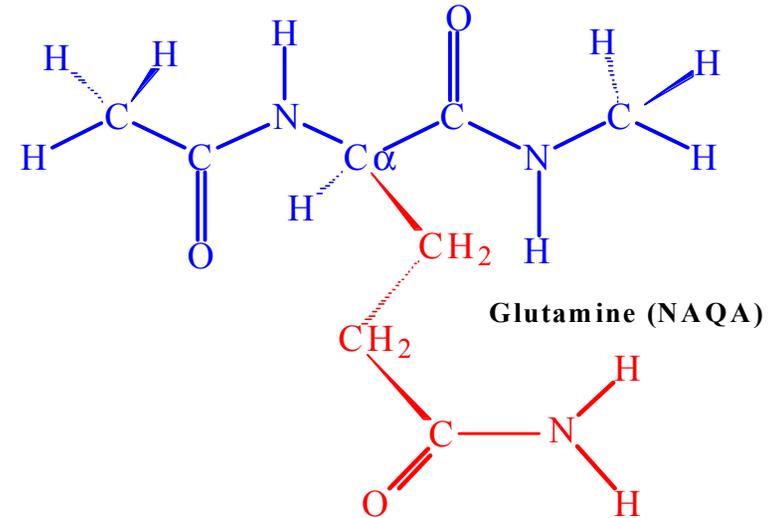
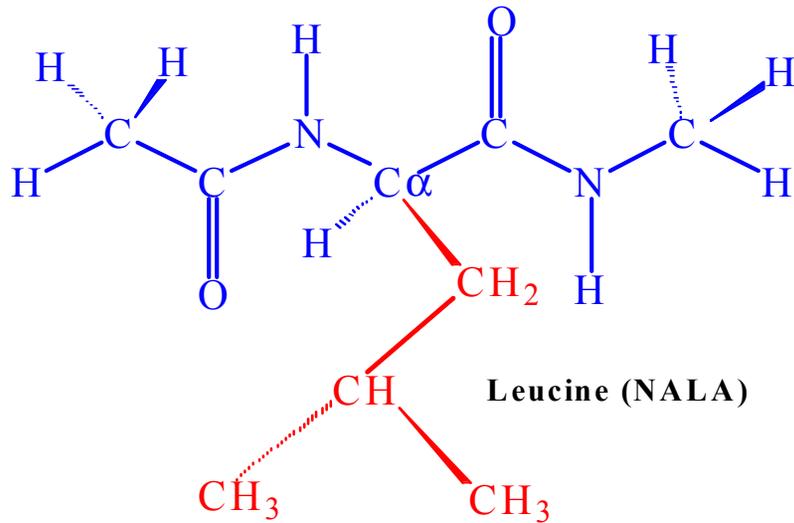


A biopolymer is distinct from a heteropolymer in one very important way
It's 3-D structure is uniquely tailored to perform a specific function



NMR, X-ray and electron crystallography solve structures slowly (1/2-3 yrs.)

The “Beads” are Chemically Complex Structures



Critical Assessment of Techniques for Protein Structure Prediction (CASP)



Large-scale experiment to assess protein structure prediction methods

It consists of three parts

the collection of targets for prediction from the experimental community

X-ray and NMR on structures about to be solved

the collection of blind predictions from the modeling community

comparative modeling

threading/fold recognition

ab initio prediction

the assessment and discussion of the results

CASP1: was held in December, 1994.

33 protein prediction targets

34 prediction groups took part submitting over 100 predictions

CASP2: was held in December, 1996 (Docking included)

42 protein prediction targets

70 prediction groups took part submitting over 900 predictions

CASP3: was held in December, 1998 (Docking dropped)

43 protein prediction targets

98 prediction groups took part submitting over 4000 predictions

Comparative Modeling



Only method that can provide models with an r.m.s.d. error lower than 2Å

All current comparative modeling methods consist of four sequential steps:

(1) identify proteins with known 3D structures related to the target sequence

search for templates with high sequence identity

(2) align them with target sequence and pick structures to use as templates

very dependent on step (1)

(3) build model for target sequence given alignment with the template structures

very dependent on step (2)

(4) model is evaluated using a variety of criteria.

Can usually detect where errors are large

Rigid body: superposition of rigid bodies from related sub-structures

Segment matching: combine database of structure segments

Satisfaction of spatial restraints: distance geometry or optimization to satisfy spatial restraints from alignment of target sequence with homologous known templates

However alignment quality is the most critical aspect of success



If sequence identity > 50% alignment is straightforward to construct

there are not many gaps, structural differences limited to loops/side-chains.

1 Å r.m.s.d. over 90% of residues

If sequence identity is between 30%-40%; alignment is not as straightforward

gaps in alignment more frequent/longer; structural differences become larger

1.5 Å r.m.s.d. for 80% of residues; rest of residues modeled with large errors

If sequence identity below 30%; comparative modeling is usually unsuccessful

Insertions longer than 8 residues cannot be modeled accurately at this time

CASP1: Problems in alignment

Some major mistakes in geometry: D-amino acids

CASP2: Some improvement in alignment accuracy

general improvement loop modeling

progress in automation

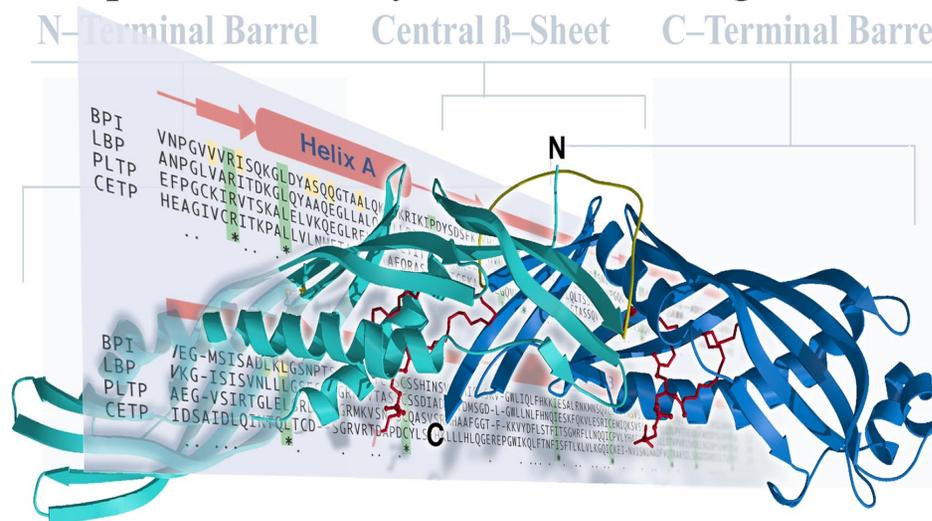
CASP3: no discernable improvement in alignment accuracy, overall prediction

blurred distinction between comparative modeling and threading

Fold Recognition, Inverse Folding, and Threading



When sequence identity is in the "twilight zone" <25-30%



Sequence Assignments to Protein Fold Topology (David Eisenberg, UCLA)

Fold Assignment: Which member of library does target sequence best match?

Inverse Fold Assignment: Find sequences in genome that match a given fold

Threading: Actual tertiary structure prediction

Results: 25% of sequences recognize their folds with high confidence

~25% of folds do not exist in library

~50% of time because need better scoring functions/alignments

100,000's of sequences/10,000's of structures (each 10^2 - 10^3 amino acids long)

Protein Fold Recognition: Threading



Uni-positional objective function

Define 1D profile of 3D tertiary structure using aa properties

Define compatibility of twenty aa's for each position in the 1-D profile

Dynamic programming: 1-D alignments of test sequence to profile: optimal alignments of subsequences extends to optimal alignments of whole; objective functions are one-dimensional $E = \sum V_i + \sum V_{\text{gap}}$

Complexity: scales as L^2 Whole human genome: 10^{13} flops

Multi-positional objective function

NP-hard if variable-length gaps and model nonlocal effects

Recursive dynamic programming, HMM's, stochastic grammars

Complexity: scales as L^3 Whole human genome: $\sim 10^{16}$ flops

CASP1: Alignments poor

All target folds not predicted by any one group

False positives a problem

CASP2: Alignments better

No clear consensus on best method/approach

False positives less of a problem; targets were easier than Casp1!

CASP3: Alignment improvements stall

Remote fold relationships poorly detected

Return of uni-positional methods: HMM's in twilight zone

Ab Initio Structure Prediction



Less reliance on knowledge-based approaches

Methods:

Secondary structure prediction

Predicting residue contacts

Conformational search techniques based on energy function

CASP1: overall much less good than comparative modeling/fold recognition

Strong movement toward fold recognition

CASP2: some improvement (still not competitive with fold recognition)

secondary structure state of the art established

PhD (~72% residues correct on average)

prediction of rough topologies of small proteins

CASP3: No formal pre-division of targets: comparative modeling, threading and ab initio.

'ab initio' methods increasingly relying on structure knowledge: 'mini-threading'

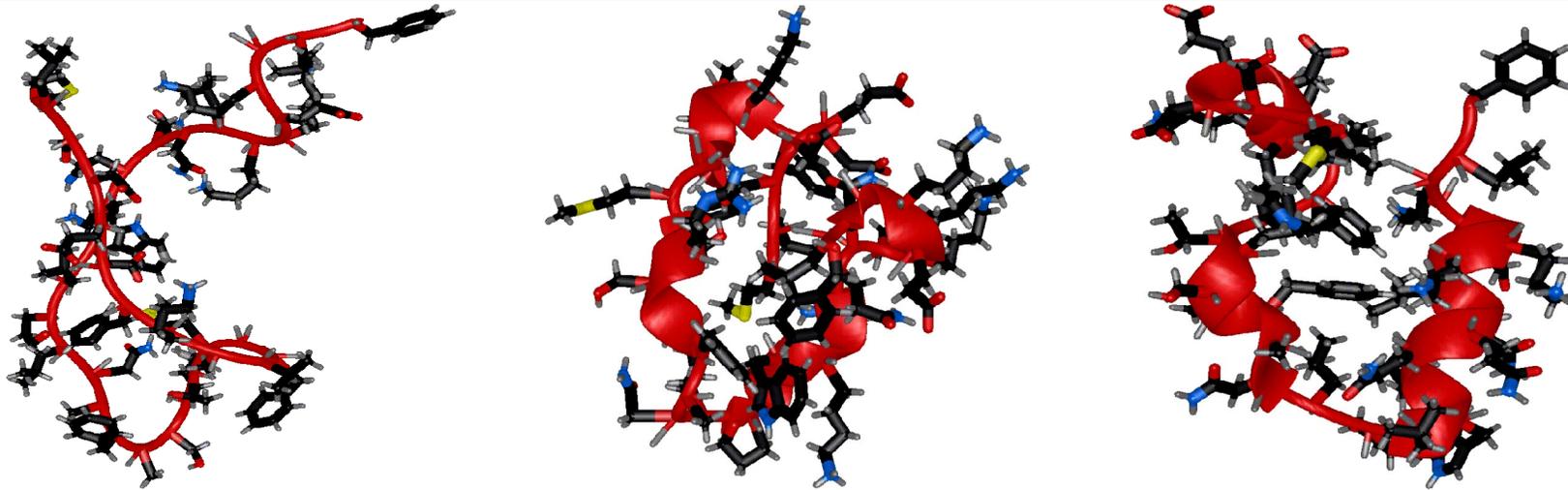
true ab initio predictions: r.m.s.d. over 10Å for 90% of predictions

Longest predicted fragment under a cut-off r.m.s.d.

80% of sequence within 4Å r.m.s.d. is a good prediction

100% of sequence within ~6Å r.m.s.d. is a good prediction

Computational Protein Folding



One microsecond simulation of a fragment of the protein, Villin. (Duan & Kollman, Science 1998)

(1) robust objective function ✓

all atom simulation with molecular water present: some structure present

(2) severe time-scale problem ✓

required 10^9 energy and force evaluations: parallelization (spatial decomposition)

(3) proper treatment of long-ranged interactions X

cut-off interactions at 8\AA , poor by known simulation standards

(4) Statistics (1 trajectory is anecdotal) X

Many trajectories required to characterize kinetics and thermodynamics

Computational Protein Folding



(1) Size-scaling bottlenecks: Depends on complexity of energy function, V

Empirical (less accurate): cN^2 ; ab initio (more accurate): CN^3 or worse ; $c \ll C$

empirical force field used

“long-ranged interactions” truncated so cM^2 scaling; $M < N$

spatial decomposition, linked lists

(2) Time-Scale of motions bottlenecks (Δt)

$$r_i(t + \Delta t) = 2r_i(t) - r_i(t - \Delta t) + \frac{f_i(t)(\Delta t)^2}{m_i 2!} + O[(\Delta t)^4] ;$$

$$v_i(t) = \frac{r_i(t + \Delta t) - r_i(t - \Delta t)}{2\Delta t} + O[(\Delta t)^3]$$

$$f_i = m_i a_i = -\nabla_i V(r_1, r_2, \dots, r_N)$$

Use timestep commensurate with fastest timescale in your system

bond vibrations: 0.01Å amplitude: 10^{-15} seconds (1fs)

Shake/Rattle bonds (2fs)

Multiple timescale algorithms (~5fs) (not used here)

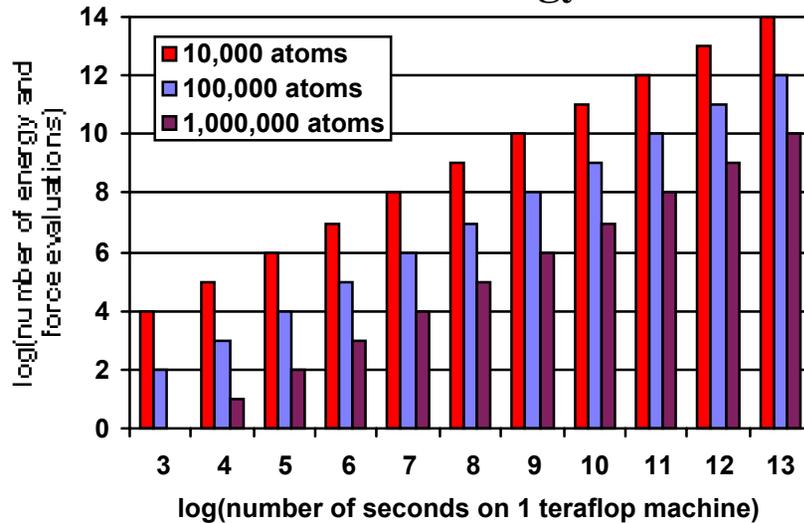
1 Microsecond simulation of Villin Headpiece in Water



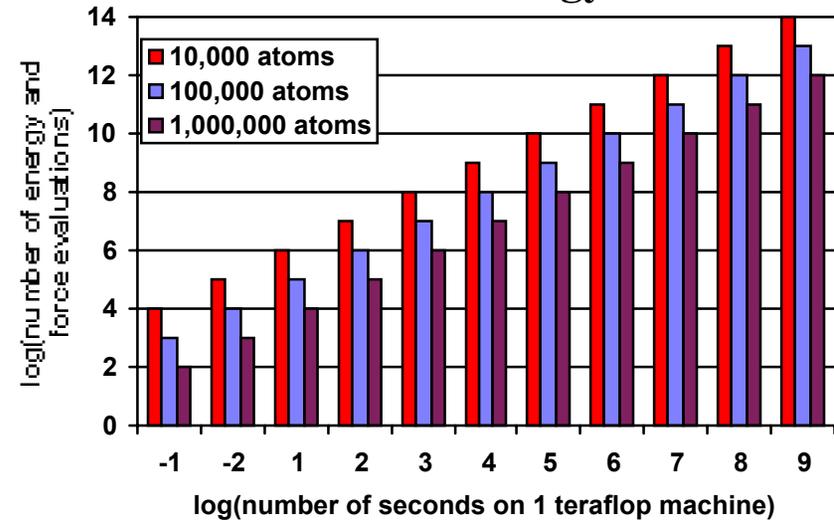
Generate 10^9 steps on 1 teraflop machine

1000 Flops per energy/force evaluation

N^2 evaluation of energy & forces



N evaluation of energy & forces



Ewald Sums:

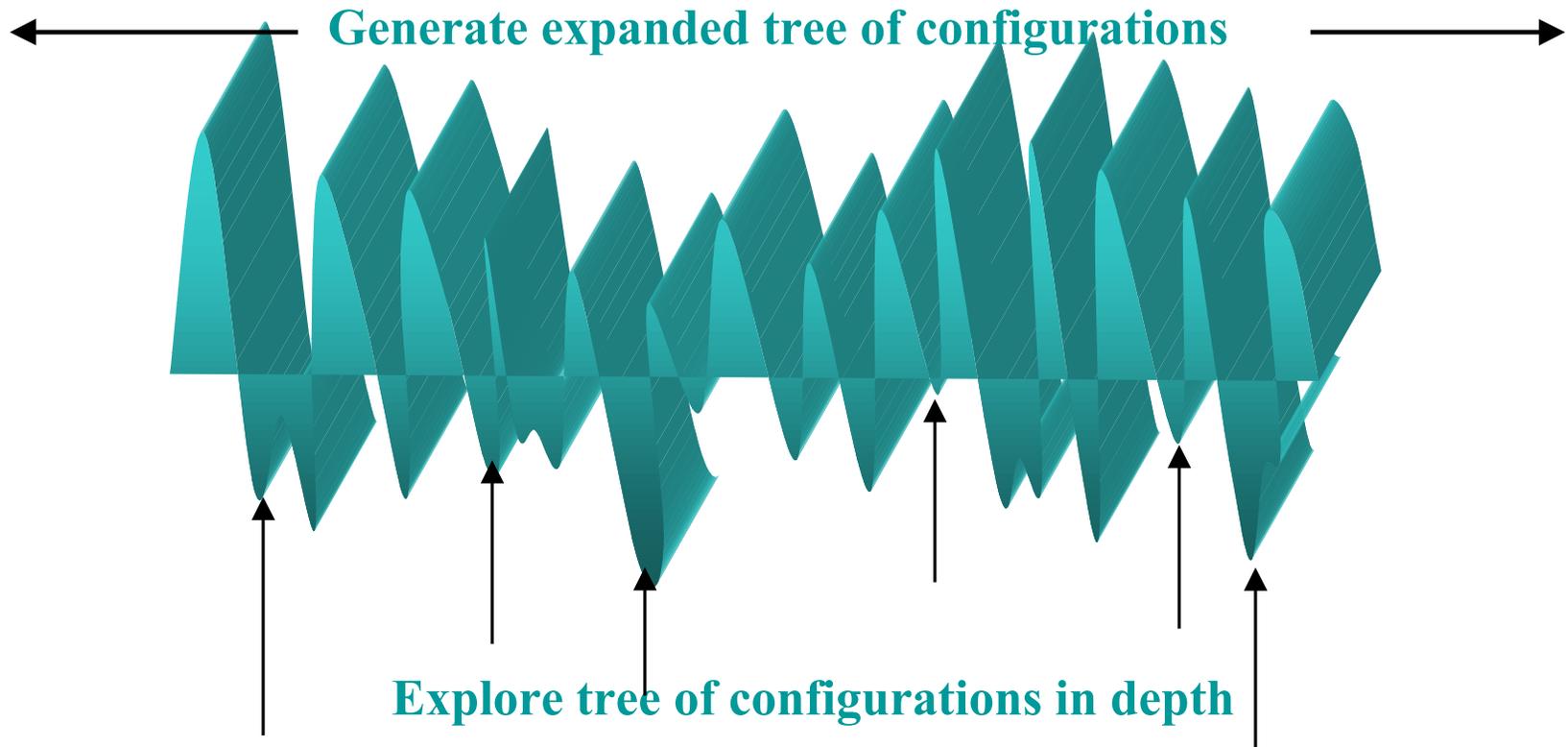
$$V_{qq} = \sum_{i>j}^N \left(\sum_{|\mathbf{n}|=0}^{\infty} q_i q_j \frac{\text{erfc}(\kappa |\mathbf{r}_{ij} + \mathbf{n}|)}{|\mathbf{r}_{ij} + \mathbf{n}|} + \frac{1}{\pi L^3} \sum_{\mathbf{k} \neq 0} q_i q_j \frac{4\pi^2}{k^2} \exp(-k^2 / 4\kappa^2) \cos(\mathbf{k} \cdot \mathbf{r}_{ij}) \right) + V_{self}$$

- Particle Mesh Ewald (N)

Spatial Decomposition in r-space; Parallelization of FFT's in k-space

- Evaluate full Ewald sum in r-space using FMM techniques

Ab Initio Protein Structure Prediction

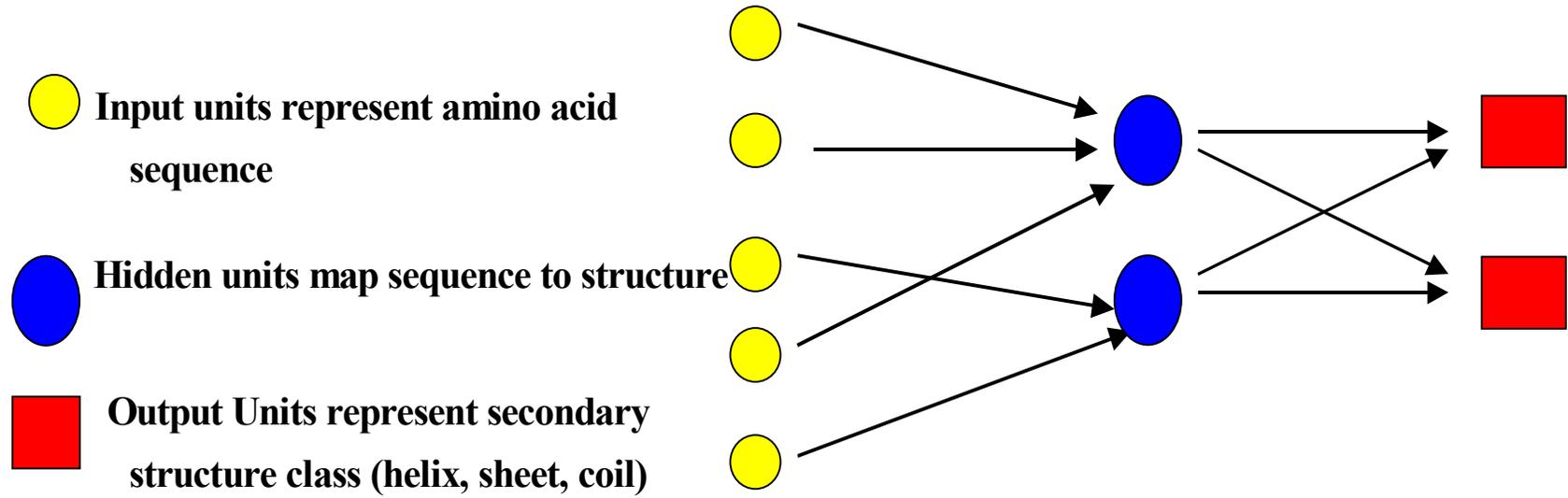


Sequence, an objective function, a search method

Tertiary Structure

- ◆ Incorporation of Constraints Predicted by Machine Learning Methods
- ◆ Global Optimization Approach to Predict Tertiary Structure
- ◆ Parallelization of Tree Search Problems
- ◆ Protein and Aqueous Solvent Energy Surface

Designed Neural Networks to Predict Protein Secondary Structure



→ Weights are optimizable variables that are trained on database of proteins

Three factors inhibit machine learning:

Adequacy of Training Database

Network Connectivity

Multiple Minima Problem in Space of Network Variables

Network Connectivity Design Yu & Head-Gordon, Phys. Rev. E (1995)

input and output representation

number of hidden neurons

weight connection patterns that detect structural features

Neural Network Results



No sequence homology through multiple alignments

Networks without design

Train

Total predicted correctly = 66%

Helix: 51% $C_\alpha=0.42$

Sheet: 38% $C_\beta=0.39$

Coil: 82% $C_c=0.36$

Test

Total predicted correctly = 62.5%

Helix: 48% $C_\alpha=0.38$

Sheet: 28% $C_\beta=0.31$

Coil: 84% $C_c=0.35$

Network with Design: Yu and Head-Gordon

Train

Total predicted correctly = 67%

Helix: 66% $C_\alpha=0.52$

Sheet: 63% $C_\beta=0.46$

Coil: 69% $C_c=0.43$

Test

Total predicted correctly = 66.5%

Helix: 64% $C_\alpha=0.48$

Sheet: 53% $C_\beta=0.43$

Coil: 73% $C_c=0.44$

Combine networks of Yu and Head-Gordon with multiple alignments

Global Optimization Algorithm: Stochastic Perturbation



Stochastic/perturbation in sub-space of dihedral angles predicted coil

- (1) Local minimization of a set of start points in sub-space
- (2) Define a critical radius

$$r_k = \left[\left(\frac{1}{\pi} \right)^{n/2} \Gamma \left(1 + \frac{n}{2} \right) \frac{V \sigma \log \rho}{\rho} \right]^{1/n}$$

a measure of whether a point is within a basis of attraction

- (3) Generate many sample points in sub-space volume, V
- (4) Evaluate r.m.s. between new sample points and minimizers of (1)
If (r.m.s. $< r_k$) ignore this sample point
- (5) Minimize sample points not in critical distance, merge into (1)

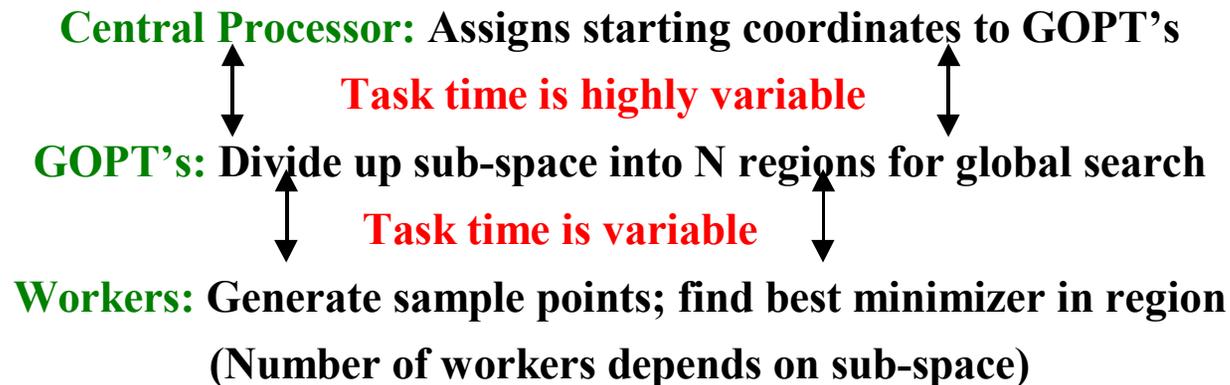
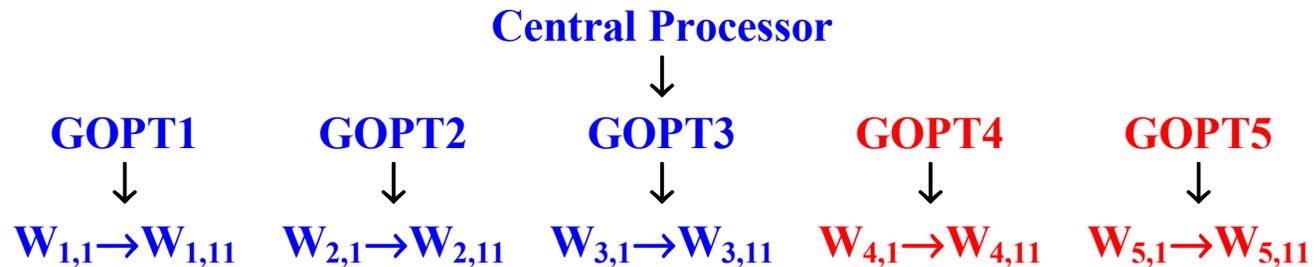
Choose new set of coil dihedral angles and repeat

Crivelli, Philip, Byrd, Eskow, Schnabel, Yu, Head-Gordon (1999). In *New Trends in Computational Methods for Large Molecular Systems*, in press.

Probabilistic theoretical guarantees of global optimum in sub-spaces
Global optimization of full space: solve series of global optimum in sub-spaces?

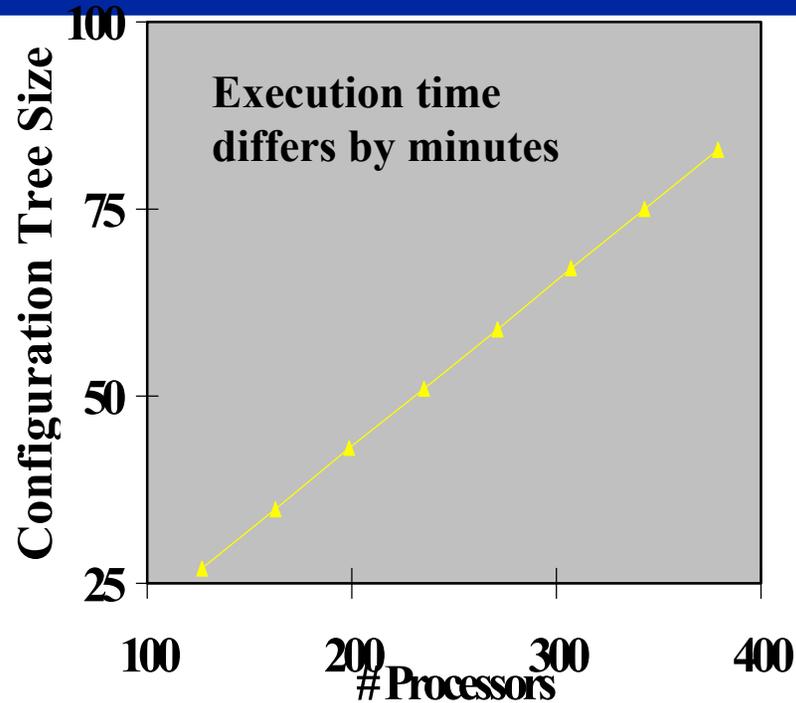
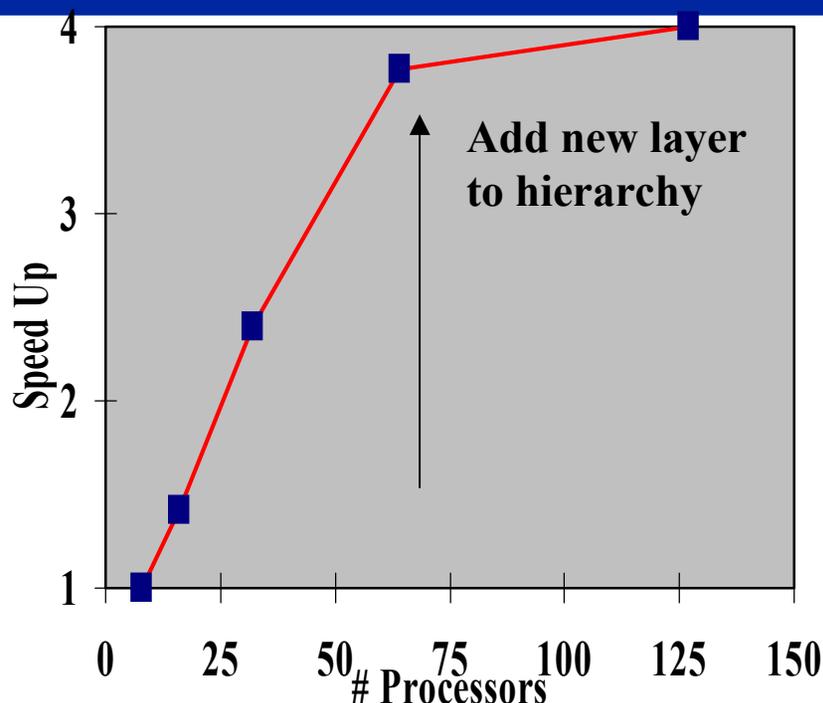
Static vs. Dynamic Load Balancing of Tasks

The work complexity to reach a minimum is highly variable



Dynamical load balancing of tasks: reassigning **GOPT/workers** to **GOPT/workers**

Static vs. Dynamic Load Balancing of Tasks



Static Number of Tree Nodes: 14 configurations

Increasing number of supervisors

Gain in efficiency of a factor of 4-8 depending on job size

Expanded Number of Tree Nodes: 14 to 83 conformations

No loss in scalability

Hierarchical/Dynamic Load Balancing

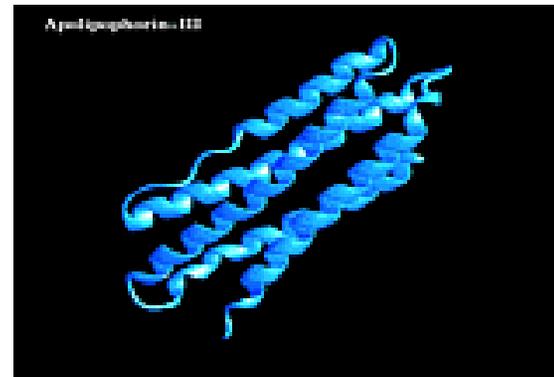
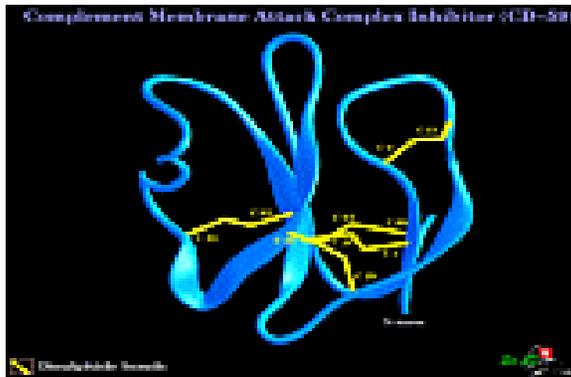
Generic to large tree search problems

Crivelli & Head-Gordon (2000). Submitted to *J. Parallel & Distributed Computing*

The Energy (Objective) Function



Empirical Protein Force Fields: AMBER, CHARMM, ECEPP "gas phase"



CATH protein classification: <http://pdb.pdb.bnl.gov/bsm/cath>

α -helical sequence/ β -sheet structure

β -sheet sequence/ α -helical structure

Energies the same! Makes energy minimization difficult!

Add penalty for exposing hydrophobic surface: favors more compact structures

$E_{\text{native folds}} < E_{\text{misfolds}}$ for a few test cases

Solvent accessible surface areas: Numerically difficult to use in optimization

We have used potentials of mean force similar to our scattering experiments

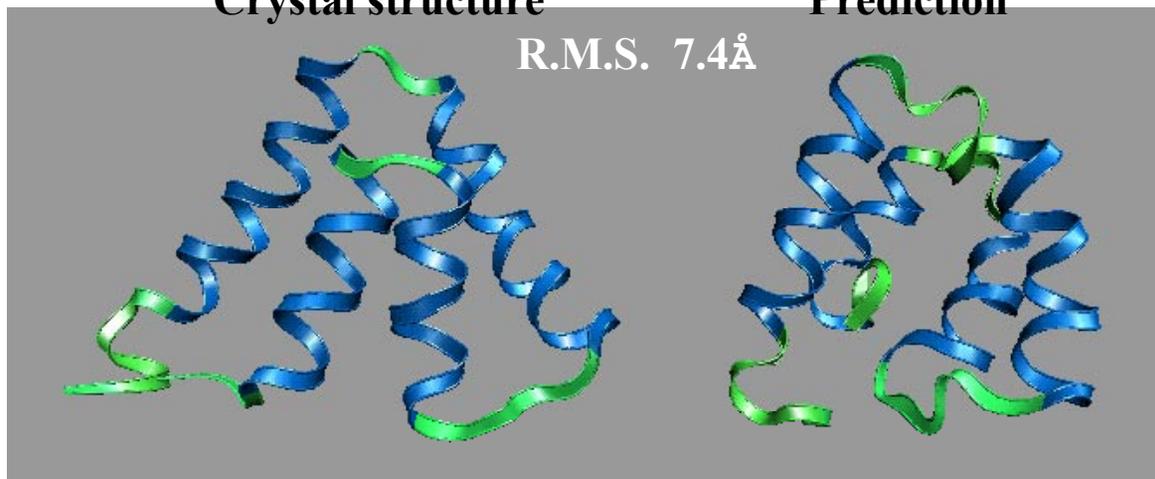
Results on α -Helical Proteins

2utg_A: 70aa α -chain of uteroglobin

Crystal structure

Prediction

R.M.S. 7.4Å



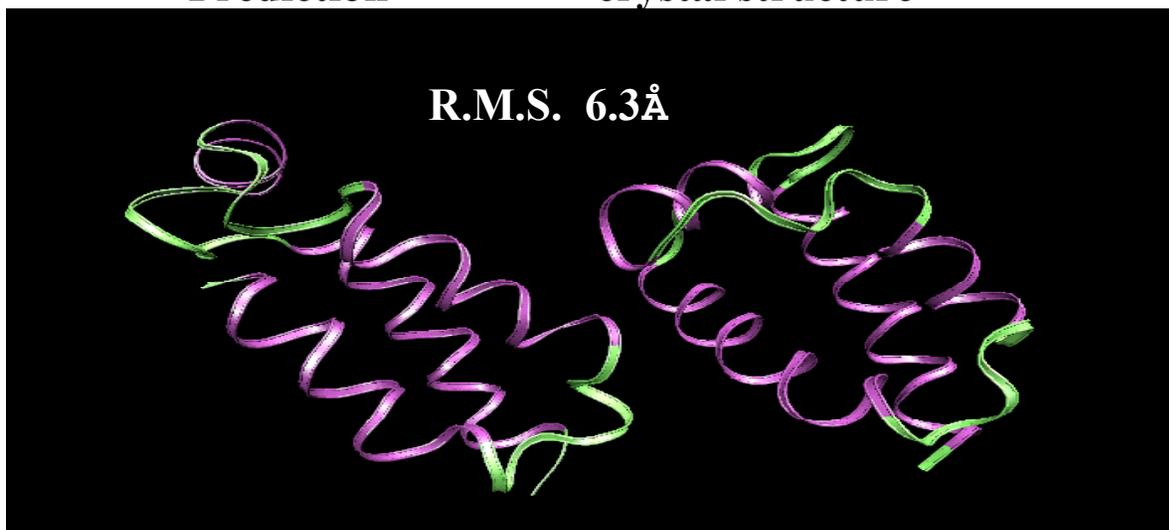
We still have not reached
crystal structure energy
yet!

1pou: 72 aa DNA binding protein

Prediction

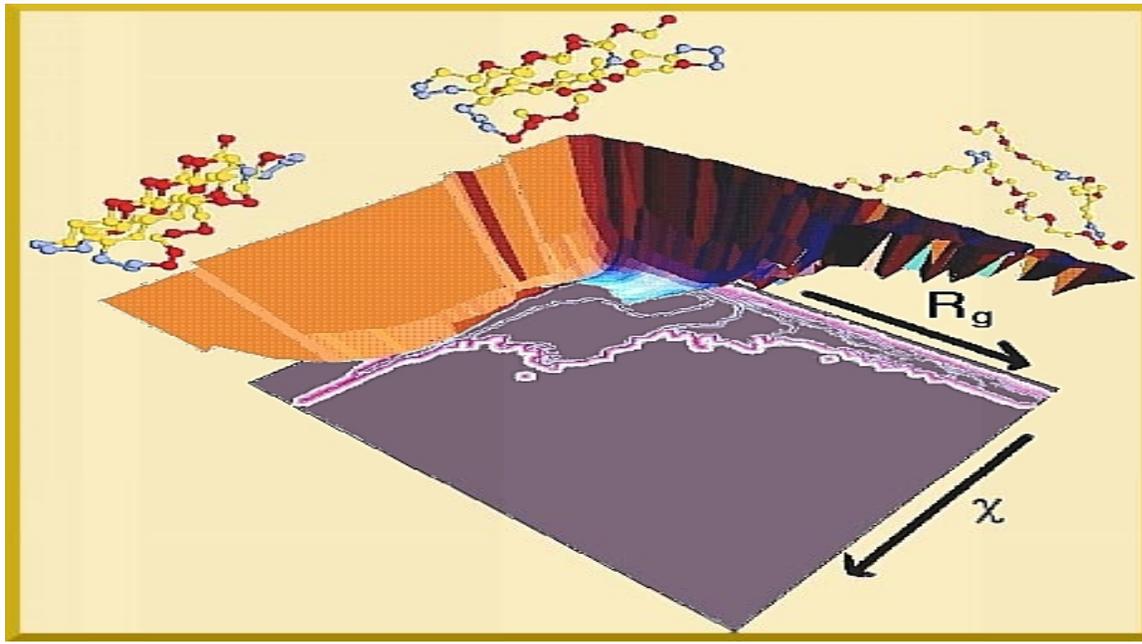
crystal structure

R.M.S. 6.3Å



As good as results for
CASP3!
Gearing up for CASP4

Simplified Models for Simulating Protein Folding for Whole Genomes



Simplifies the “real” energy surface topology sufficiently that you can do:

(1) Statistics ✓

Can do many trajectories to converge kinetics and thermodynamics

(2) severe time-scale problem ✓

characterize full folding pathway: mechanism, kinetics, thermodynamics

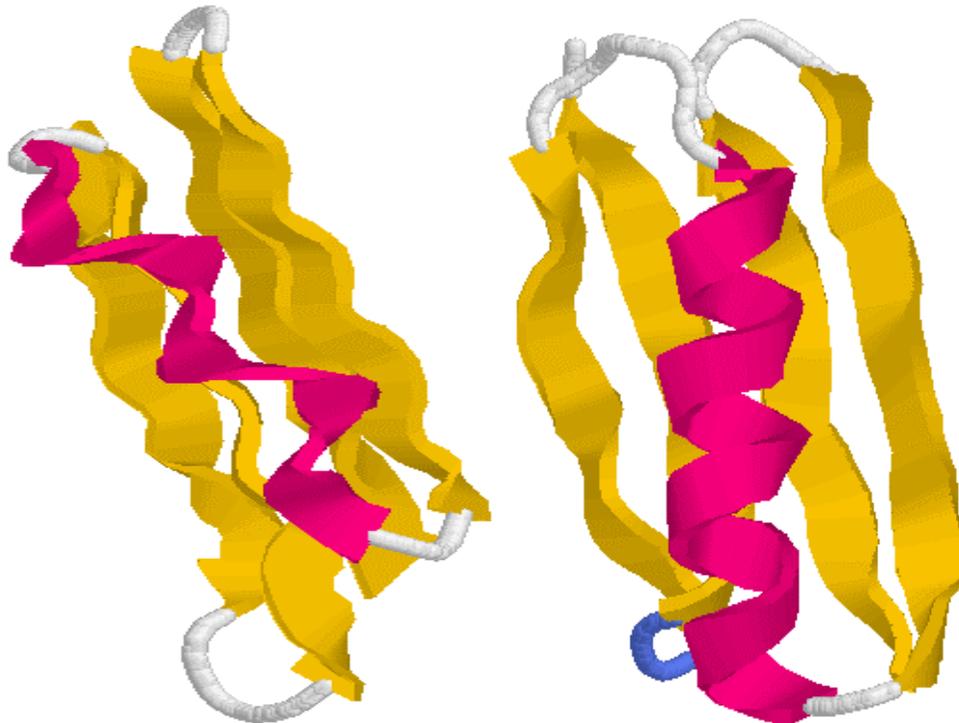
(3) proper treatment of long-ranged interactions ✓?

all interactions are evaluated; no explicit electrostatics

(4) robust objective function?

good comparison to experiments

Protein Folding Model of IgG-Binding Protein's L & G



Model of Protein G/L

Protein L

Protein chain is 56-residues

Sequence is specified as:

- ◆ hydrophobic (*B*), hydrophilic (*L*), or neutral (*N*) beads
- ◆ secondary structure dihedrals

Benefit of separation of 1° and 2°:

- ◆ Role of non-local (bead-bead) and local (dihedral) interactions
- ◆ Changing secondary structure propensity: mutation studies
- ◆ increases complexity of model: more than three flavors

$$H = \sum_{\text{angles}} \frac{1}{2} k_{\theta} (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} \{ A [1 + \cos \phi] + B [1 - \cos \phi] + C [1 + \cos 3\phi] \\
 + D [1 + \cos (\phi + \pi/4)] \} + \sum_{i, j \geq i+3} 4 \epsilon_H S_1 \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - S_2 \left(\frac{\sigma}{r_{ij}} \right)^6 \right]$$

J. M. Sorenson & T. Head-Gordon (2000). Accepted to RECOMB'2000, Tokyo

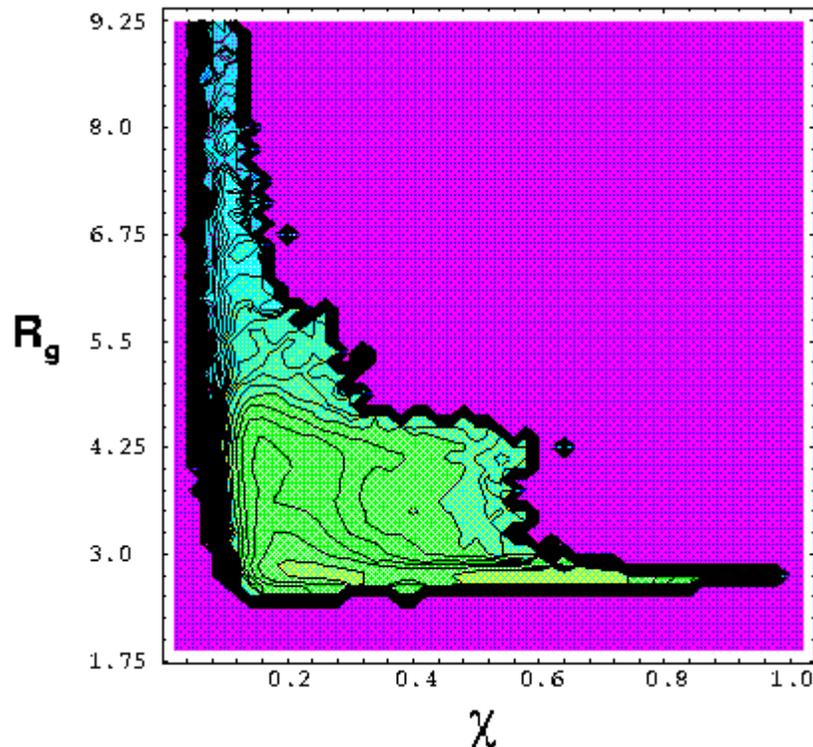
Characterization of Free Energy Landscape: Multiple Multi-Histogram Method



Six-dimensional histograms over **energy** and **five order parameters**

R_g , χ , χ_H , $\chi_{\beta 1}$, and $\chi_{\beta 2}$

$$\chi = \frac{1}{M} \sum_{i,j \geq i+4}^K \theta(\epsilon - |r_{ij} - r_{ij}^{Nat}|)$$



Collapse-concomitant-with-folding scenario

More accessible states closer to the diagonal

Non-compact structure, partly native structure

Collapse transition is at only a slightly higher temperature than the folding transition

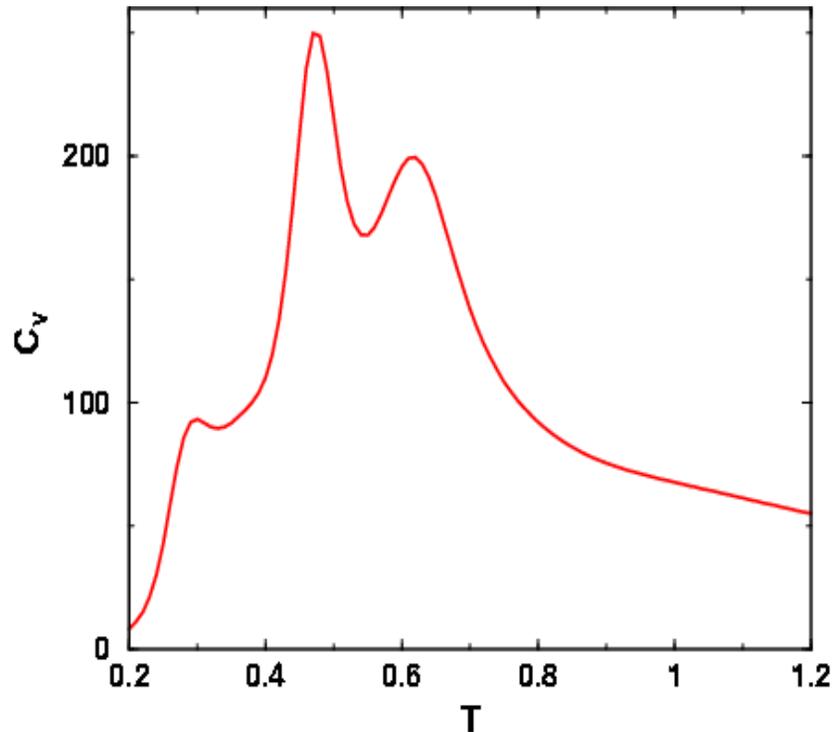
Recent time-resolved SAXS experiments on protein L (Doniach, Baker 1999):

Chain collapse is rate-limiting, occurring closely in time to native state formation

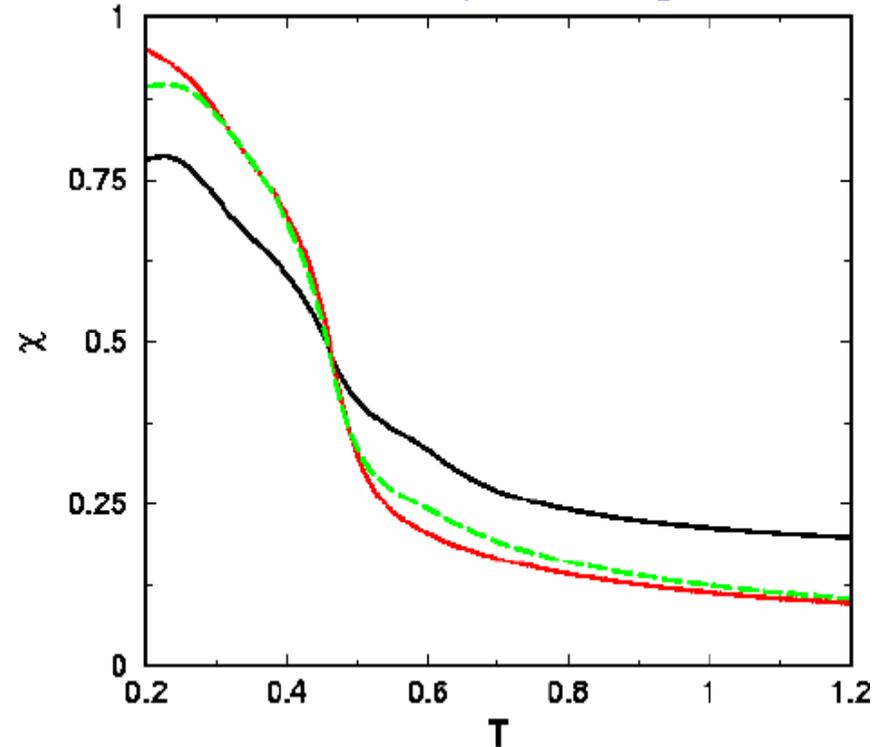
Results for α/β Protein Fold: Thermodynamics



Heat capacity vs. Temperature



2° Structure similarity v.s. Temperature



- ◆ $T=0.62$: weak transition to form some helical structure, without beta hairpin structures.
- ◆ $T=0.46$: all three secondary structure order parameters show native-like structure
- ◆ Major peak in the heat capacity curve corresponds to the folding transition

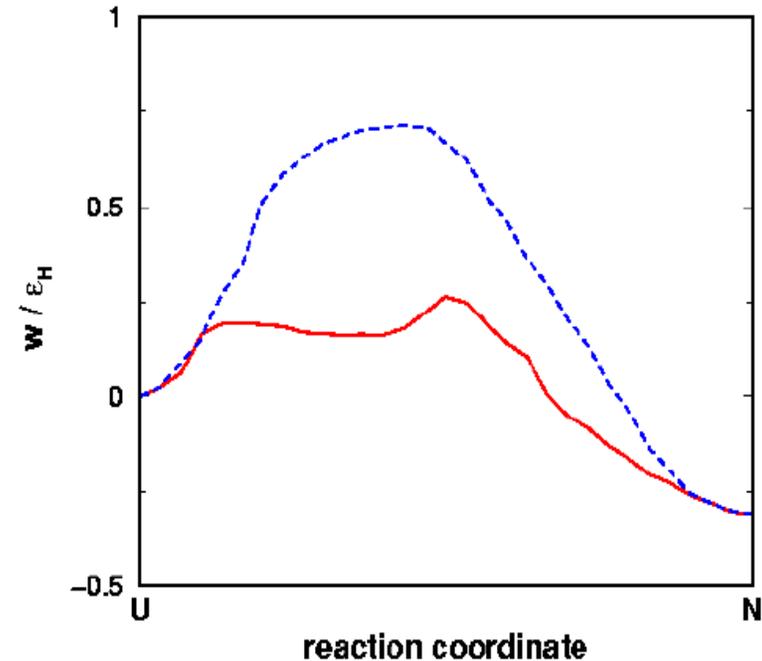
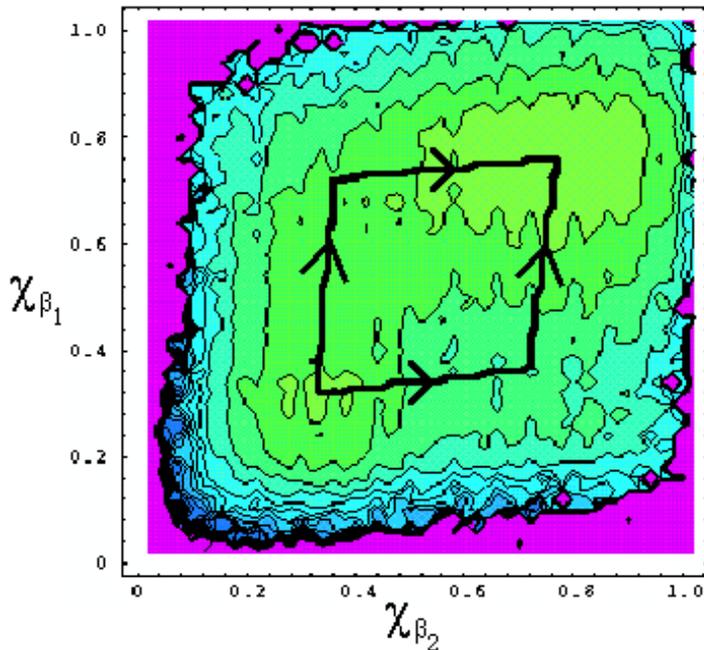
Strong support for the high cooperativity of the folding process in this model

Free Energy Surface is Asymmetric to β -hairpin Formation



Consider two kinetic mechanisms

- (1) β -hairpin #1 and helix first, then form β -hairpin #2: **Intermediate**
- (2) β -hairpin #2 and helix first, then fold β -hairpin #1: **No Intermediate**



While Protein L and G have nearly identical tertiary structure:

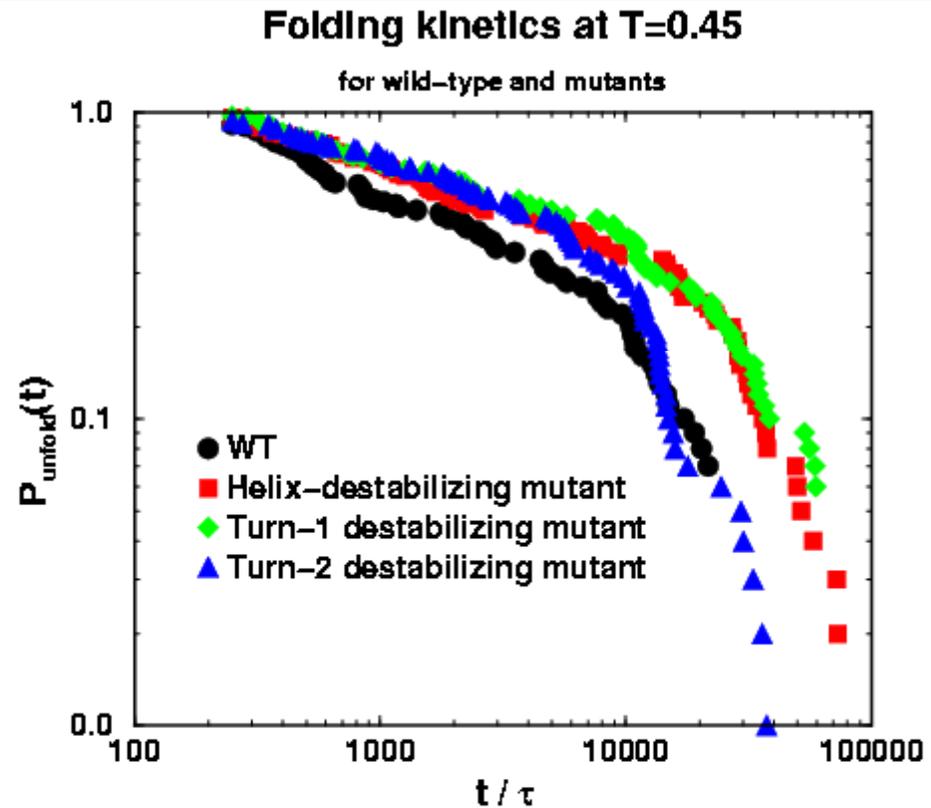
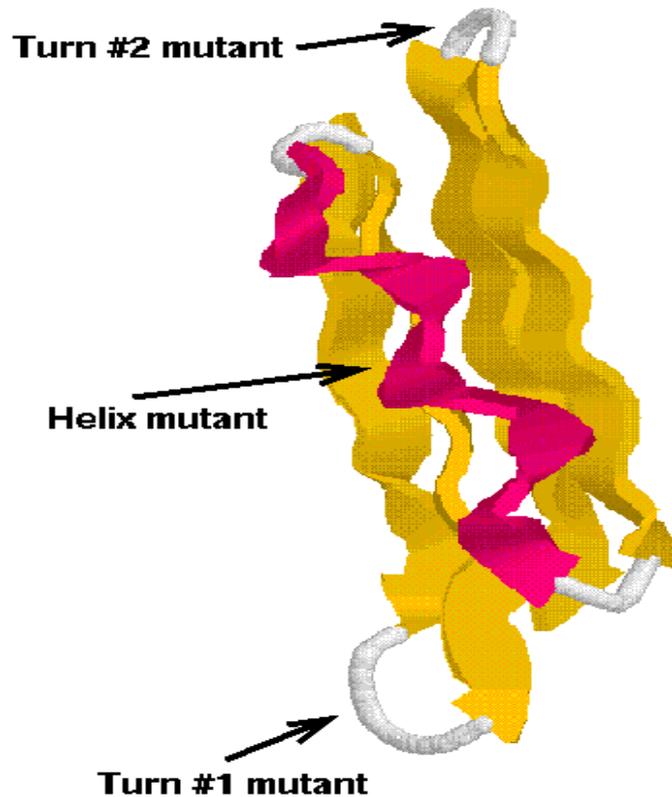
Protein G has high energy intermediate along the folding pathway

Park, O'Neil, Roder, Biochemistry (1997)

Protein L appears to be purely two-state with a single barrier

Yi, Baker, Protein Sci. (1996)

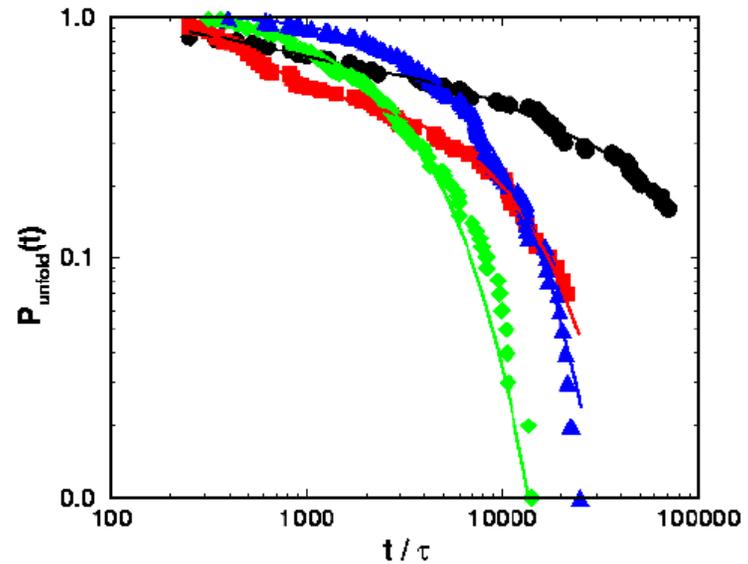
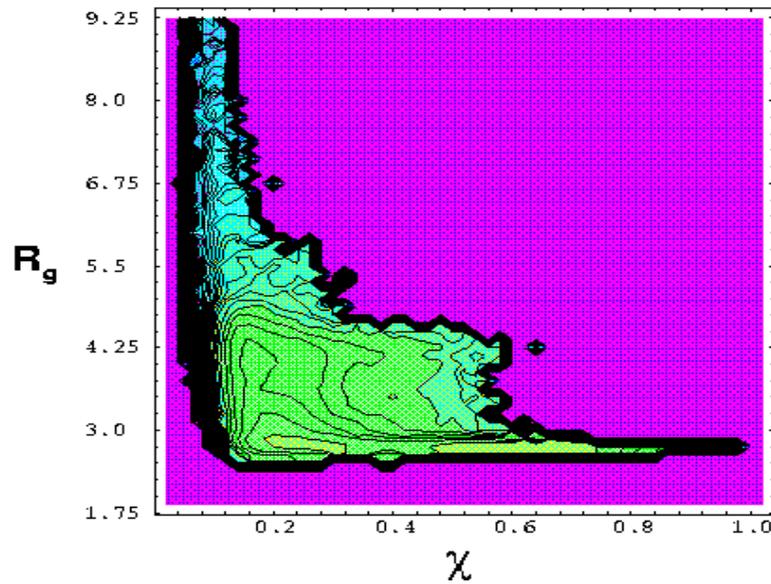
Mutations on α/β Protein consistent with Mutation Studies of Protein L



helix-destabilized and turn#1-destabilized mutants fold noticeably slower
consistent with their formation as intermediate in dominant folding pathway
turn#2-destabilized mutants fold as fast as wildtype
turn#1 is not destabilized

Gu, Kim, Baker, J. Mol. Biol. (1997); Baker et al., J. Mol. Biol. (1998)

Computational Complexity of Simplified Models for Protein Folding



Thermodynamics of the folding process are characterized using

multi-histogram method: complexity increases with multiple order parameters

constant-temperature Langevin simulations

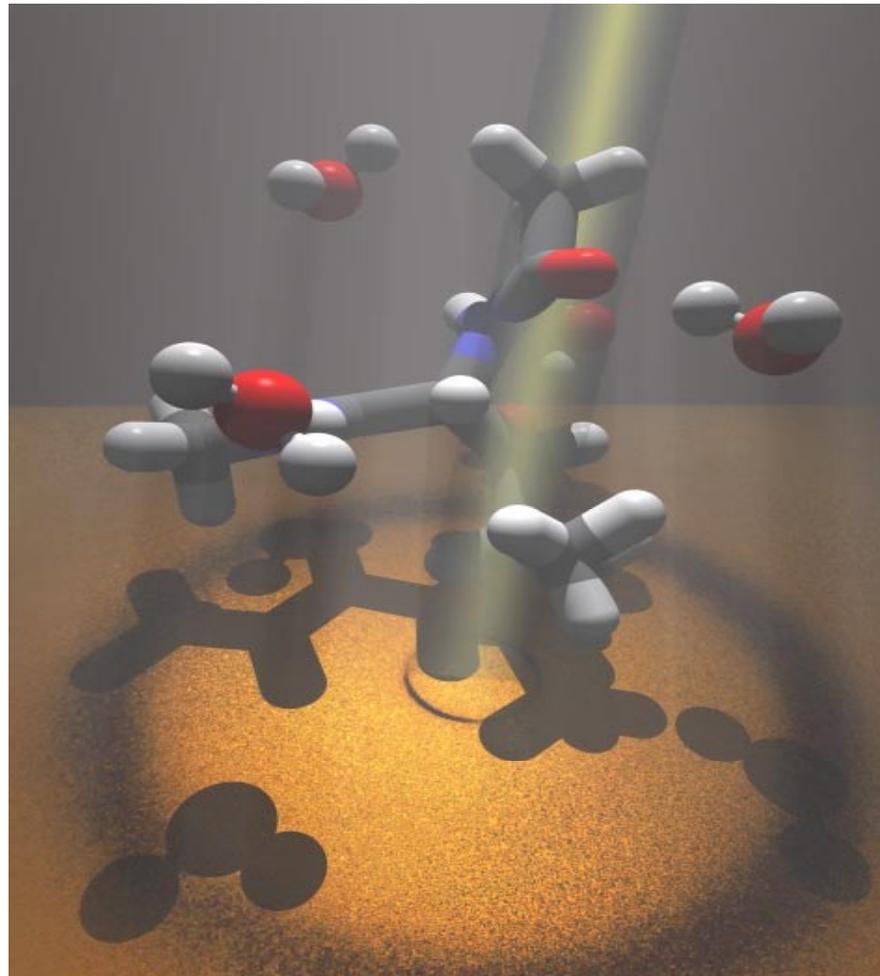
Folding kinetics are characterized by tabulating

mean-first passage times, and temperature scans

One week using two Compaq/Dec EV10000 (~50 specfp95) per protein sequence

100,000 sequences for Human Genome; Ample mutational study data

Determining the Role of Hydration Forces in the Folding of Model Proteins



Scattered x-ray beam and resulting image pattern from N-acetyl-leucine-methylamide in water. Cover of Feature Article in J. Phys. Chem. B.

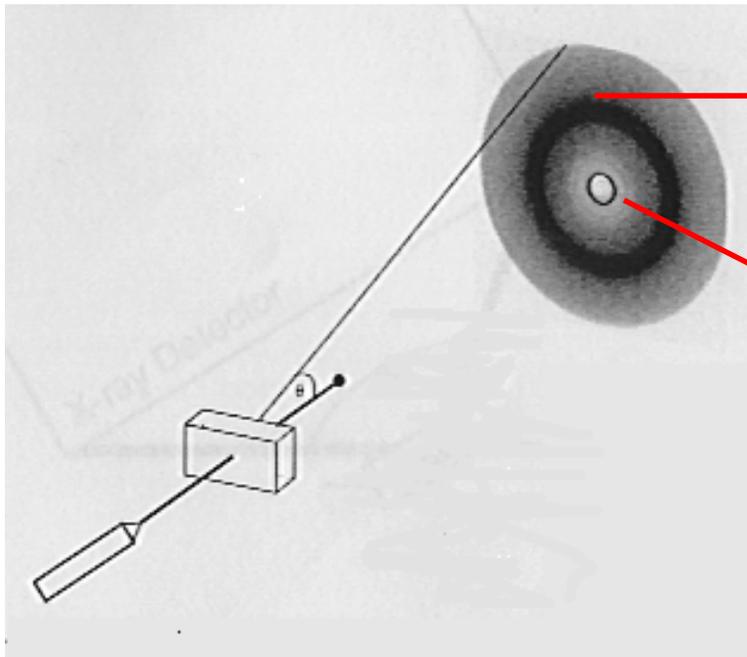
Sorenson, Hura, Soper, Head-Gordon (1999), 103 5413-5426

Concentrated solutions of N-acetyl-Leucine-methylamide

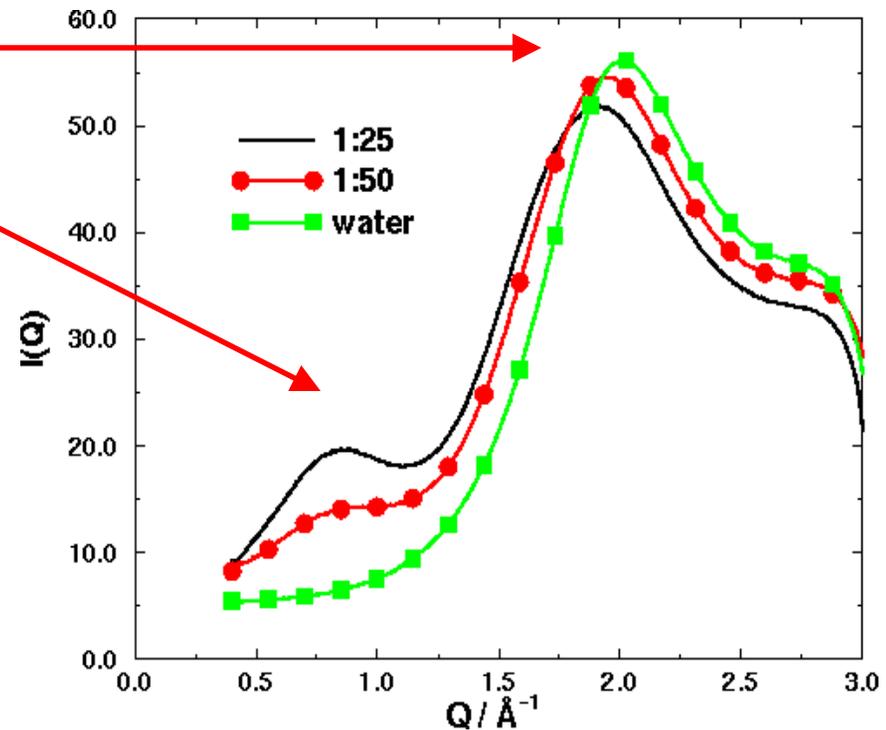
Model of Late Protein Folding Events



X-ray Solution Scattering



N-acetyl-leucine-methylamide in Water



at a concentration of 1:50 a new feature appears at $Q \sim 0.8 \text{ \AA}^{-1}$

develops into peak at same effective Bragg spacing maximum concentration of 1:25

represents a stable solute-solute configuration

Hura, Sorenson, Glaeser & Head-Gordon (2000). In [Perspectives in Drug Discovery & Design](#), V. 17, 1-12.

Molecular Dynamics Simulation of Excess Scattering Intensity.



$$I_{\text{solution}}(Q) = I_{\text{solute-solute}}(Q) + I_{\text{solute-water}}(Q) + I_{\text{water-water}}(Q) + I_{\text{intramolecular}}(Q)$$

$$I(Q) = \sum_A \sum_B c_A c_B b_A b_B H_{AB}(Q), \quad H_{AB}(Q) = 4\pi\rho \int_0^\infty r^2 [g_{AB}(r) - 1] \frac{\sin(Qr)}{Qr} dr,$$

b_A : neutron scattering length for A $g(r)$: radial distribution function
 c_A : atomic fraction of A ρ : total number of atoms per unit volume
 Q : momentum transfer

Simulation:

AMBER 95: for solutes (solute-water interactions parameterized with TIP3P)

SPC: for water (like TIP3P but slightly better water structure)

Water-water: $g_{OO}(r)$, $g_{OH}(r)$, $g_{HH}(r)$ **Water-solute**: $g_{OX}(r)$ and $g_{HX}(r)$

Simulation Details:

NVT ensemble (Nosé-Hoover chain of thermostats)

1.5fs timestep, Rattle, Ewald sums

150ps equilibration, 300ps statistics

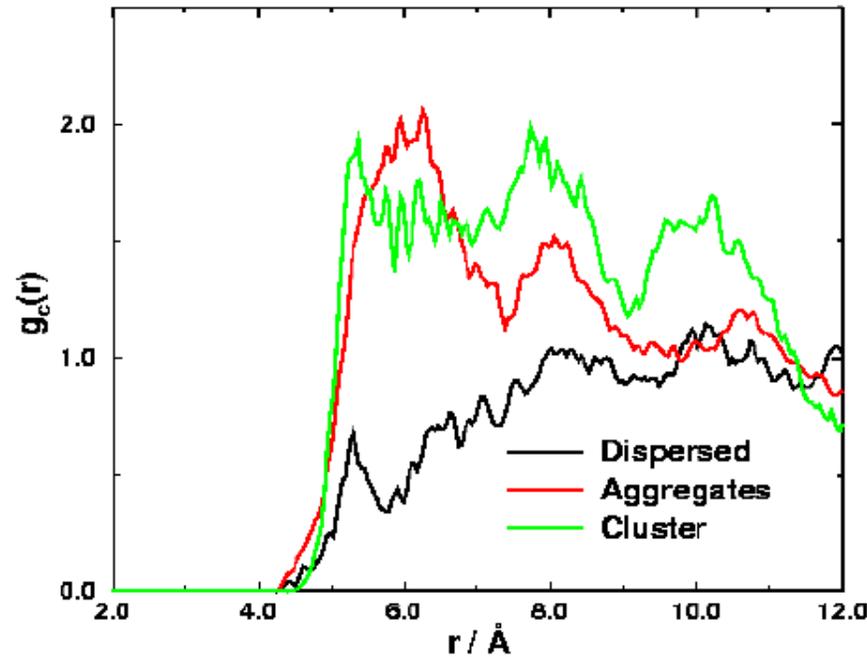
Interpretation of New Diffraction Feature at 0.8\AA^{-1} for NALMA in Water



What is $I_{\text{solute-solute}}(Q) \longrightarrow g_C(r)$?

Centers radial distribution function

for 1:24 NALMA configurations



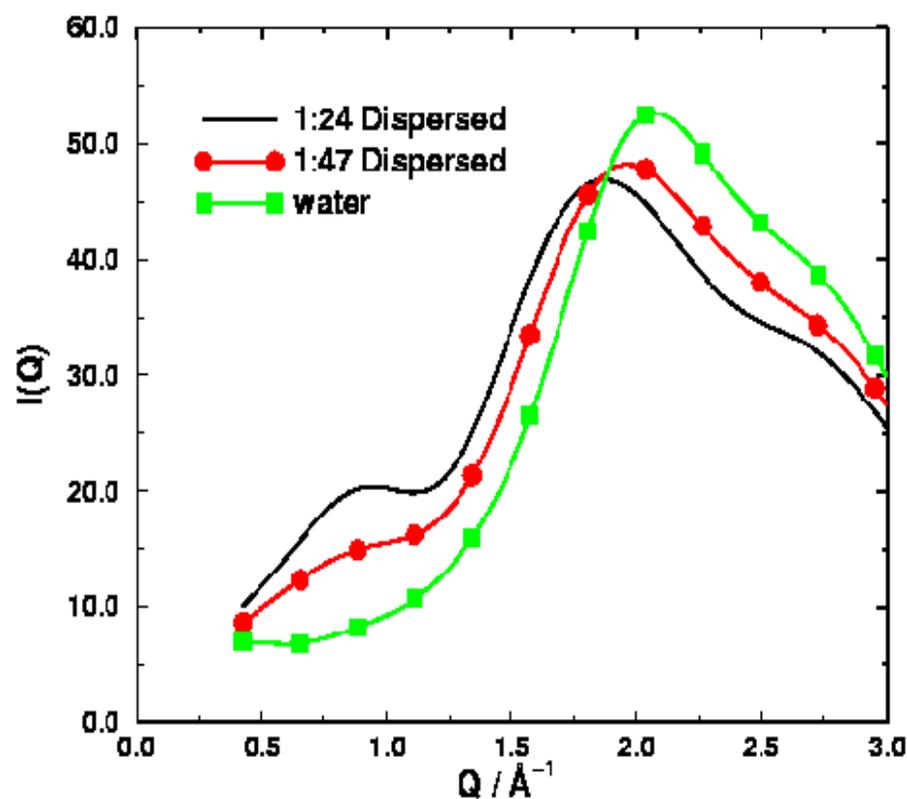
Consider qualitatively different solute distribution function

Dispersed

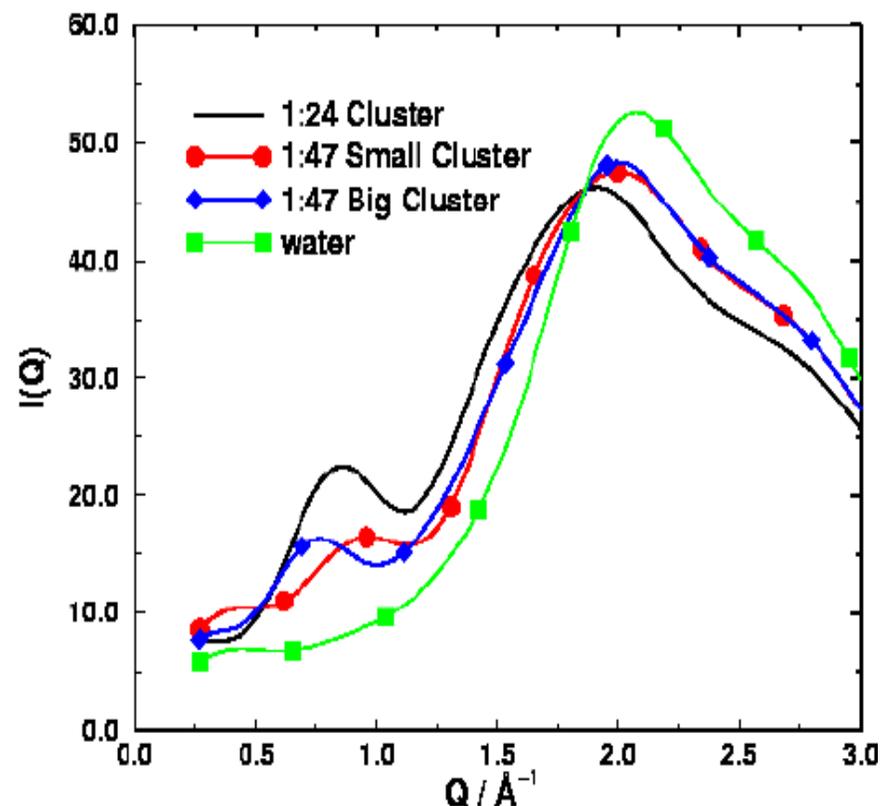
Small clusters

Single large cluster

Dispersed and small NALMA aggregates are preferred



Dispersed/Small Clusters



Large Clusters

What is right: Hydrophilic/hydrophobic character of the protein backbone/side chains
local side chain and backbone conformational entropy costs

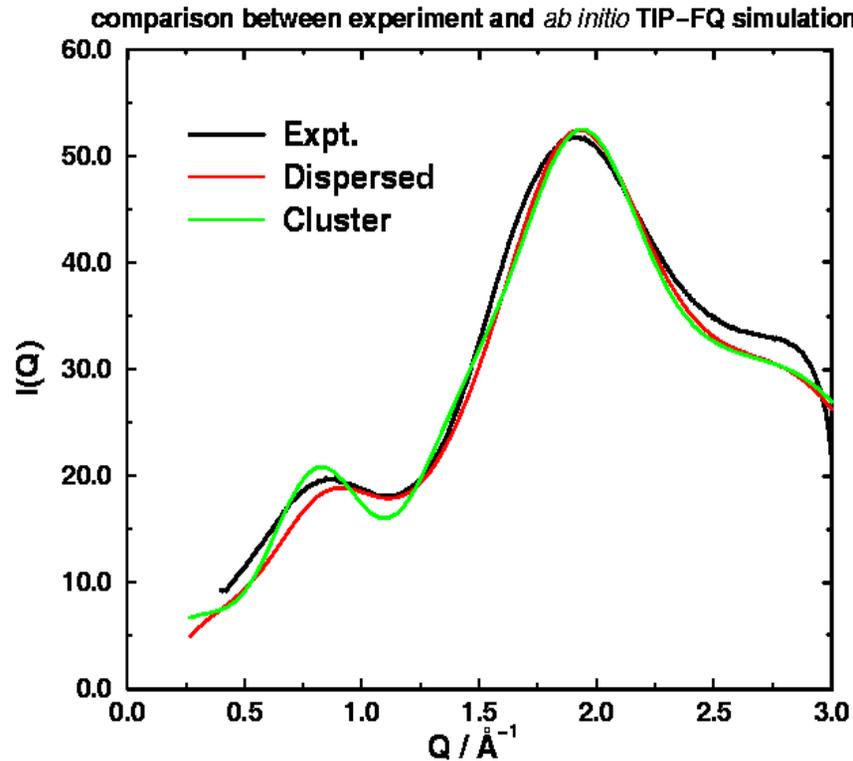
What is questionable: demixing entropy (monomers) vs. conformational entropy (chain)

result should be extensible to real proteins which are never this hydrophobic.

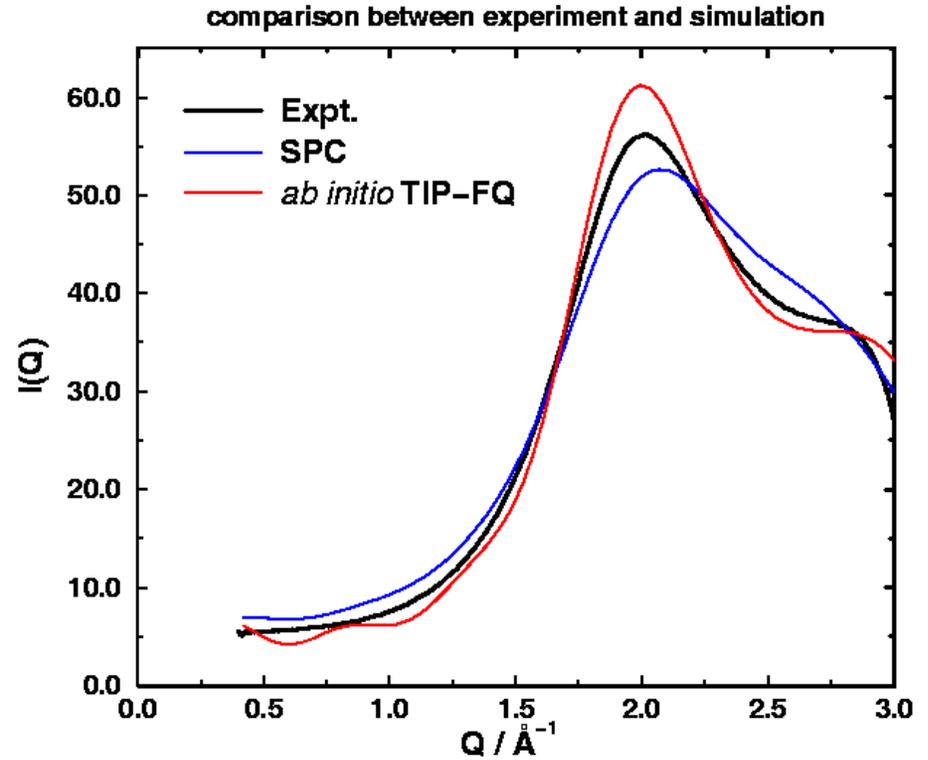
Quantitative Agreement between Experiment and Simulation



X-ray Scattering for 1:25 NALMA



Pure Water Scattering



Polarizable TIP4P model (Rick & Berne, JCP 1994)

polarizability essential for reproducing experimental observable?

However, scattering experiments over the last ~30 years are still inadequate

what is $g_{OO}(r)$?

The Role of Water for Life in Precarious Circumstances

AAAS, 2000



John A. Baross, University of Washington

Thermophilic microbes from black smoker chimneys

Results from recent joint expedition of U. Washington and American Museum of Natural History

Image from <http://www.amnhonline.org/expeditions>

How do you explain growth and function at extremes of temperature and pressure?

Amino acid composition of total protein extracted from bacterial cells only unusually high levels of glutamic acid

Are protein sequences/structures different between mesophiles and thermophiles?

Sequence correlations with increased stability were not evident (Baross 1998)

Topology differences negligible, but smaller loops, tighter core (Dansen et al., 1998)

More sensitive statistical analysis?

***Maintain stability but remain flexible (functional) in face of environmental extremes?**

Do physical hydration forces change with temperature and/or pressure? (Baross 1998)

Acknowledgements for THG Research



Silvia Crivelli, Physical Biosciences and NERSC Divisions, LBNL

**Betty Eskow, Richard Byrd, Bobby Schnabel, Dept. Computer Science,
U. Colorado**

Jon M. Sorenson, NSF Graduate Fellow, Dept. Chemistry UCB

Greg Hura, Graduate Group in Biophysics, UCB

Alan K. Soper, Rutherford Appleton Laboratory, UK

**Alexander Pertsemidis, Dept. of Biochemistry, U. Texas Southwestern Medical
Center**

**Robert M. Glaeser, Mol. & Cell Biology, UCB and Life Sciences Division,
LBNL**

Funding Sources:

AFOSR, DOE (MICS), DOE/LDRD (LBNL), NIH, NERSC for cycles