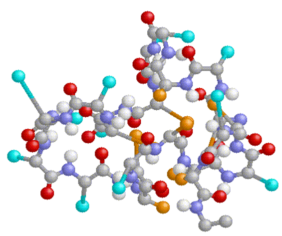
Protein Structure & Energy Landscape Dependence on Sequence using a Continuous Energy Function



A.T. Phillips
United States Naval Academy & SDSC

J.B. Rosen
University of California, San Diego & SDSC

K.A. Dill
University of California, San Francisco

Outline

- Formulation of the Protein Folding Problem
- CGU Algorithm
- Computational Platforms & Performance
- Computational Results
    - Global Minimum Conformations
    - Energy Landscapes
- Interpretation of Results
- Effect of Sequence on Structure
- Folding Dynamics

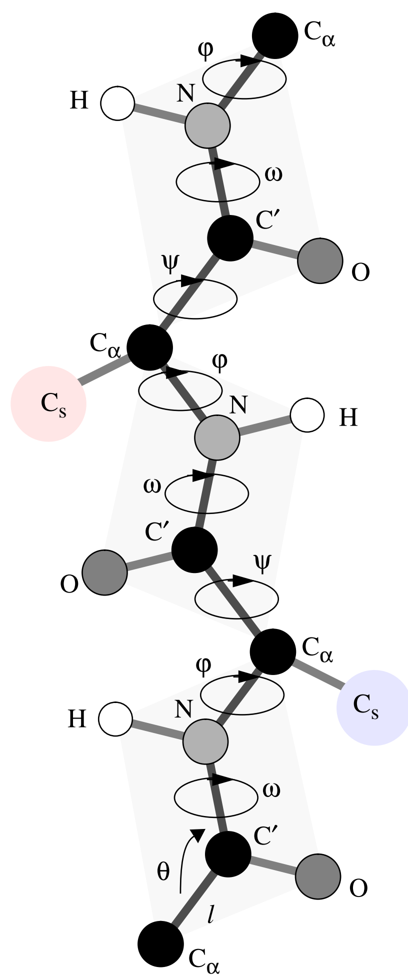
The Protein Folding Problem

- Given a known "primary" sequence of residues, predict its native, or folded, state in 3-dimensional space

Assumptions

1. The native state of the molecular structure corresponds to the global (or near global) minimum of a potential energy function.
2. Conformations are defined by internal molecular coordinates: backbone torsion angles (φ/ψ).
3. The chain of monomers consist of two types: H (hydrophobic) and P (polar/hydrophilic).
4. H-H monomer pairs are attractive.
5. All monomer pairs exhibit steric repulsive forces.

The Polypeptide Chain Model



Modeling the Potential Energy

- Recent success by independent research groups (Dill at UCSF, Rose at JHU) has shown that the dominant forces in folding are:

  1. Steric repulsion (aka excluded volume)
  2. Hydrophobic-Hydrophobic attraction
  3. Hydrogen bond formation
  4. Specific torsion angle preference

- Such an energy model becomes:

$$F(\phi) = E_{ex} + E_{hp} + E_{h} + E_{\varphi\psi}$$

The Sun/Thomas/Dill Potential Energy Function

- The model potential function is

$$F(\phi) = E_{ex} + E_{hp} + E_h + E_{\varphi\psi}$$

where:

$$1. E_{ex} = \sum_{ij} \frac{C_1}{1.0 + exp\left(\frac{d_{ij} - d_{eff}}{d_w}\right)}$$

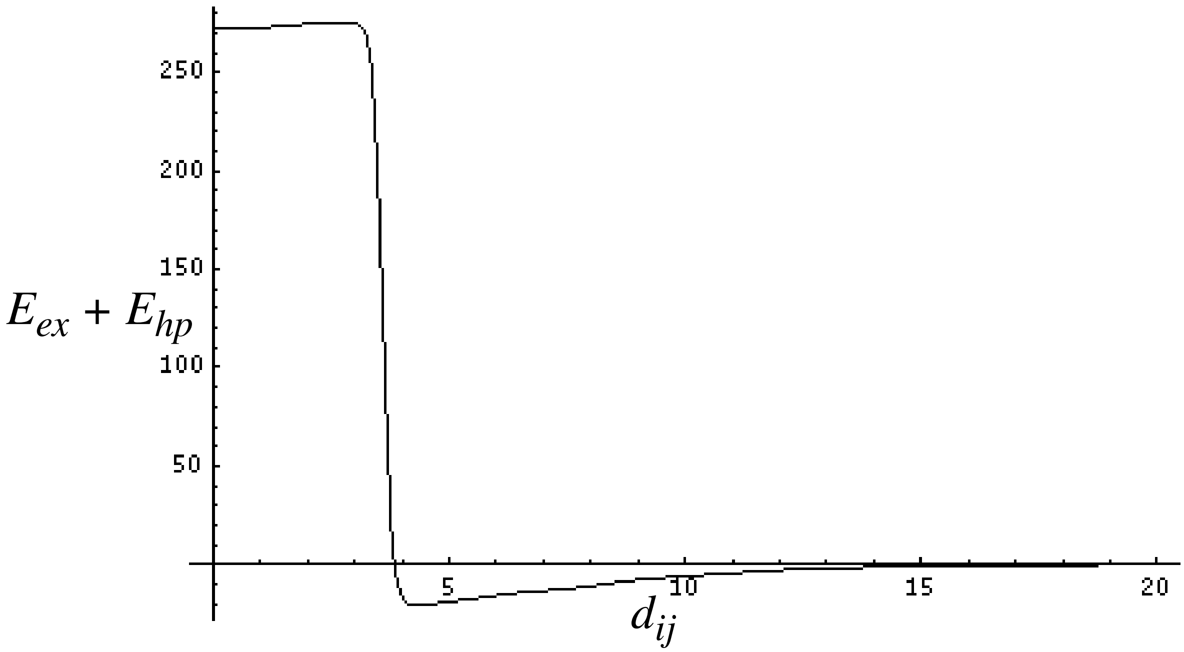
$$2. E_{hp} = \sum_{|i-j|>2} (-\varepsilon_{ij}) \frac{C_2}{1.0 + exp\left(\frac{d_{ij} - d_0}{d_t}\right)}$$

$$3. E_h = \sum_{|i-j|>3} \frac{C_3 q_1 q_2}{4\pi\varepsilon_0 D d_{ij}}$$

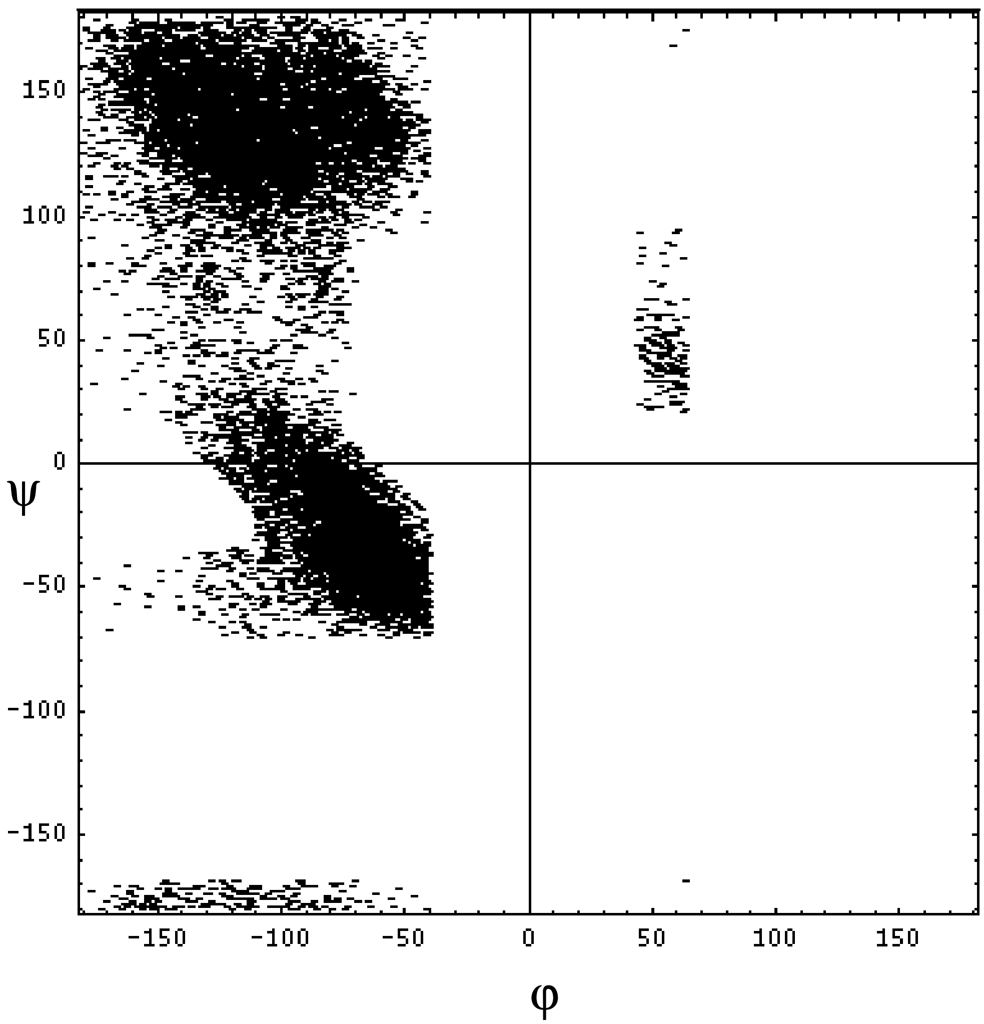
4. $E_{\varphi\psi}$ represents the preference for specific φ/ψ pairs, as shown via a Ramachandran map.

A.T. Phillips
J.B. Rosen
K.A. Dill

$E_{ex} + E_{hp}$ Energy Terms for H-H Pairs



The Ramachandran Map for All Residues Except GLY and PRO



■
A.T. Phillips
J.B. Rosen
K.A. Dill

Constructing $E_{\varphi\psi}$ to Approximate the Ramachandran Data

- Require that $\boldsymbol{E}_{\varphi\psi}$ satisfy

$$E_{\varphi\psi} = \begin{cases} 0 \ \text{if}(\varphi, \psi) \in R_i \ \text{ for some } i \\ \beta \qquad\qquad \text{otherwise} \end{cases}$$

- Represent each "ellipsoidal" region R_i by a quadratic function $q_i(\varphi,\psi)$ which satisfies the conditions:

$$q_i(\varphi,\psi) < 0 \text{ in the interior of } R_i$$

$$q_i(\varphi,\psi) = 0 \text{ on the boundary of } R_i$$

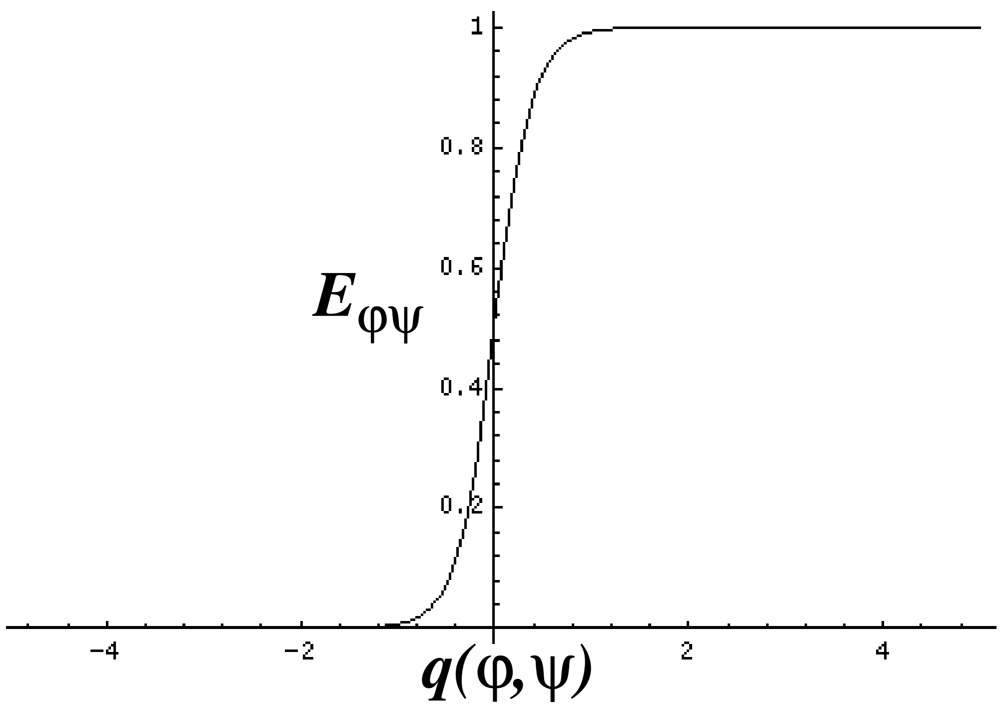
$$q_i(\varphi,\psi) > 0 \text{ in the exterior of } R_i$$

- Define the sigmoidal penalty term $\boldsymbol{E}_{\varphi\psi}$ as

$$E_{\varphi\psi} = \frac{\beta}{1.0 + \sum_{i=1}^{p} exp(-\gamma_i q_i(\varphi, \psi))}$$

■
A.T. Phillips
J.B. Rosen
K.A. Dill

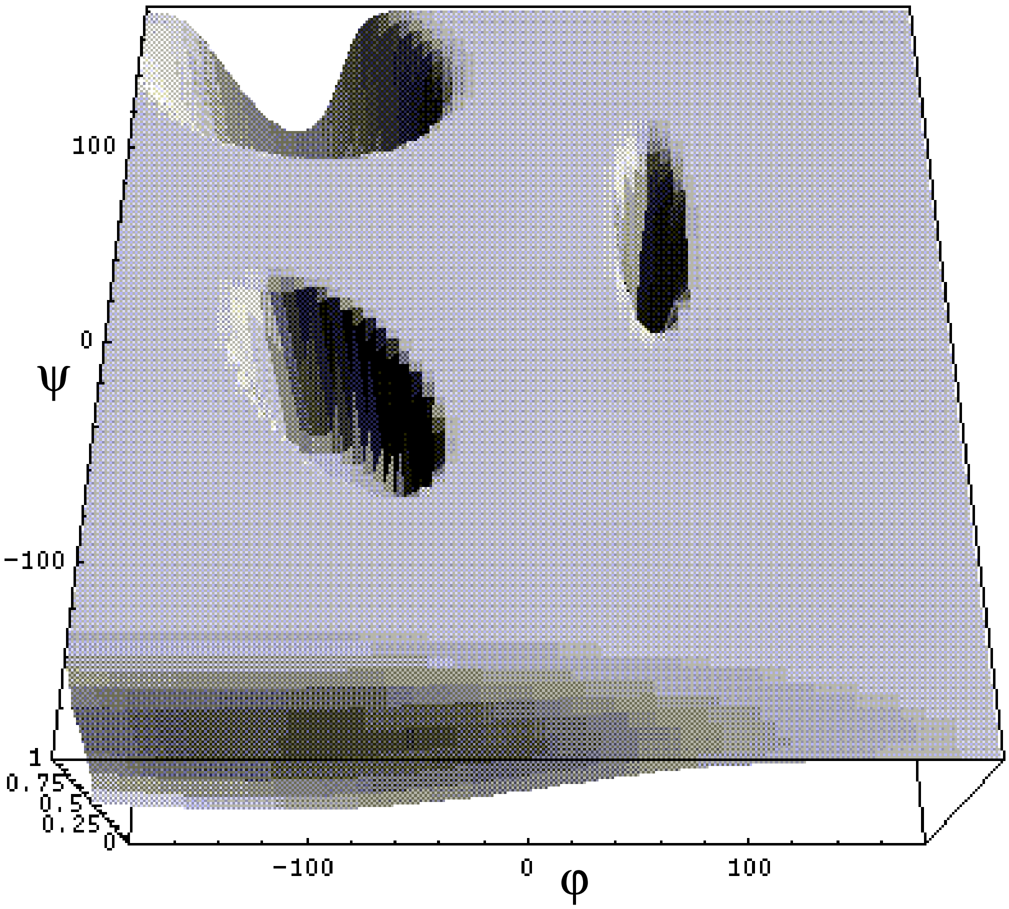
The Sigmoidal Energy Term $E_{\varphi\psi}$ with $\beta = 1$ and $\gamma = 5$



- $\beta = 1$ and $\gamma = 25$ work well for computation.

A.T. Phillips
J.B. Rosen
K.A. Dill

The Sigmoidal Approximation to the Ramachandran Map for All Residues Except GLY and PRO



- Used to implement $\boldsymbol{E}_{\varphi\psi}$ for all residues except GLY and PRO.

■
A.T. Phillips
J.B. Rosen
K.A. Dill

Molecular Conformation with Additional Distance Geometry Constraints

- Information on distances ($\boldsymbol{d_{ij}}$) between specified pairs of atoms in a molecule may be known ($\boldsymbol{r_{ij}}$) :

$$d_{ij} = r_{ij}, \text{ for } (i,j) \in S.$$

- This information can be used to improve the molecular conformation calculation.

- Add distance terms to the energy function $F(\phi)$:

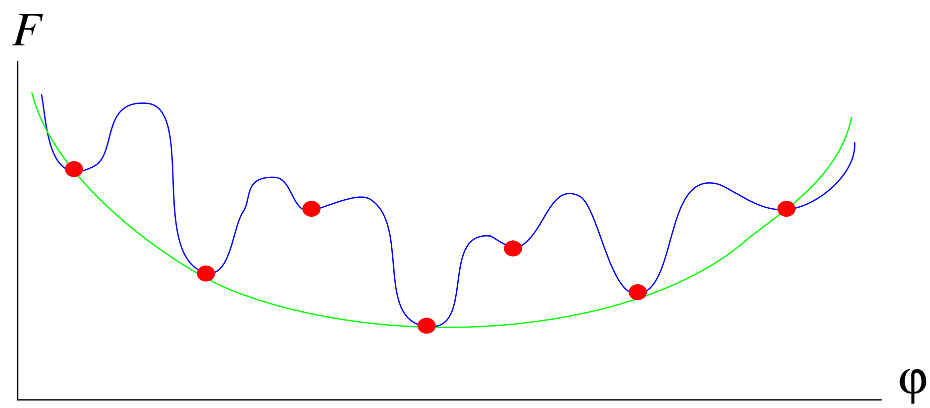
$$F_d(\phi) = F(\phi) + K_d \sum_{i,\, j \in S} (r_{ij}^2 - d_{ij}^2)^2$$

and compute the global minimum of $F_d(\phi)$.

A.T. Phillips
J.B. Rosen
K.A. Dill

Convex Global Underestimator

- Attempt to use a "global underestimating function" to localize the search in the region of the global minimum.

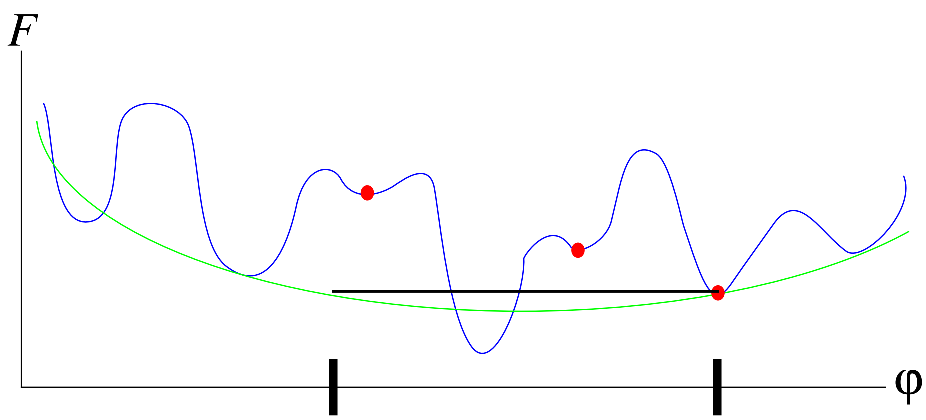


Convex Global Underestimator in One Dimension

- Fits all known local minima with a function which underestimates all points, but differs from them by the smallest possible amount (minimizes the L_1 norm).

Convex Global Underestimator (cont)

- The bounds of the hypercube Hϕ are also used to limit the "search region" around the predicted global minimum.

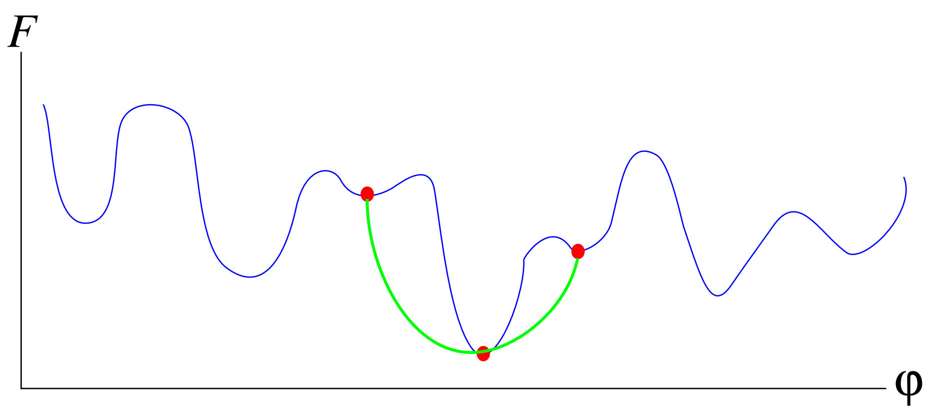


Convex Global Underestimator in One Dimension

■
A.T. Phillips
J.B. Rosen
K.A. Dill

Convex Global Underestimator (cont)

- The new more "localized" search region is explored and another convex global underestimator is computed with corresponding predicted global minimum.



Convex Global Underestimator in One Dimension

■
A.T. Phillips
J.B. Rosen
K.A. Dill

Defining the Underestimating Function

- Assume all bond lengths (l) and bond angles (θ) are fixed.
- Given k local minimizers (conformations) $\phi^{(j)}$, for j=1,...,k, determine the coefficients of the function $\Psi(\phi)$ so that:

$$\delta_j = F(\phi^{(j)}) - \Psi(\phi^{(j)}) \geq 0$$

for j=1,...,k, and where $\sum_{j=1}^{k} \delta_j$ is minimized.

■
A.T. Phillips
J.B. Rosen
K.A. Dill

Defining the Underestimating Function (cont.)

- The underestimating function

$$\Psi(\phi) = c_0 + \sum_{i=1}^{n} \left(c_i \phi_i + \frac{1}{2} d_i \phi_i^2 \right)$$

consists of linear term, c_i, and quadratic term, d_i.

- Convexity is guaranteed by requiring that $d_i \geq 0$ for i=1,...,n.

- Note that the minimum of this function is easily computed:

$\phi_i = -c_i / d_i$ for i=1,...,n.

Algorithm

1. Compute $k \geq$ 2n+1 distinct local minima $\phi^{(j)}$, for j=1,...,k, of the function $F(\phi)$.

2. Compute the convex quadratic underestimator function

$$\Psi(\phi) = c_0 + \sum_{i=1}^{n} \left(c_i \phi_i + \frac{1}{2} d_i \phi_i^2 \right)$$

by solving the linear program

$$\min_{y_1, y_2, y_3} \quad f^T y_1 - f^T e_k$$

$$\text{subject to} \quad \begin{bmatrix} \Phi & I'^T_n & -I'^T_n \\ \Omega & \underline{D} & -\bar{D} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \Phi e_k \\ \Omega e_k \end{bmatrix}, \quad y_1, y_2, y_3 \geq 0$$

3. Compute the predicted global minimum point $\phi_{\min}$ given by $(\phi_{\min})_i = -c_i/d_i$, i=1,...,n, with corresponding function value $\Psi_{\min}$ given by $\Psi_{\min} = c_0 - \sum_{i=1}^{n} \frac{c_i^2}{(2d_i)}$.

Algorithm (cont.)

4. If $\phi_{min} = \phi^*$, where $\phi^* = \mathrm{argmin}\{F(\phi^{(j)}), j=1,2,... \}$ is the best local minimum found so far, then stop and report ϕ^* as the approximate global minimum conformation.

5. Reduce the volume of the hyperrectangle $H\phi$ over which the new configurations will be produced, and remove all columns from Φ and Ω which correspond to the conformations which are excluded from $H\phi$.

6. Use ϕ_{min} as an initial starting point around which additional local minima $\phi^{(j)}$ of $F(\phi)$ (restricted to $H\phi$) are generated. Add these new local minimum conformations as columns to the matrices Φ and Ω.

7. Return to step 2.

■
A.T. Phillips
J.B. Rosen
K.A. Dill

Convergence Properties

- If the CGU underestimates the global minimum of $\boldsymbol{F}(\phi)$ at every iteration, then finite convergence to the global minimum can be guaranteed using a branch and bound method.

- Even if it fails to underestimate at some iterations, it may still give finite convergence to the global minimum.

Computation of Local Minima

- QN unconstrained minimization using BFGS updates.
- Major fraction (99%) of total computation time is used for finding local minima
- Local minimizations are easily performed in parallel --> “embarrassingly parallel”.

Computational Issues

- The algorithm is implemented in C using the MPI message passing system.
- All local minimizations are performed in parallel using all available processors.
- All other steps are performed sequentially on a single designated “master” CPU.
- Uses a “master/slave” SPMD paradigm.

Computational Platforms

- Computational tests have been conducted on:

  1. Cray T3D at SDSC using as many as 32 processors.
  2. Network of 12 Sun SparcStations and 7 SGI Indys at USNA.
  3. Dec "Alpha Farm" at SDSC using 8 processors.
  4. Intel Paragon at SDSC using as many as 64 processors.

Computational Complexity

- < $O(n^4)$ increases in time (average case)
  - Number of local minima required for construction of underestimator: $O(n)$.
  - Number of "major" iterations: $O(1)$ (always < 10).
  - Time per local minimization: < $O(n^3)$.
- $O(2^n)$ increases in time (worst case)

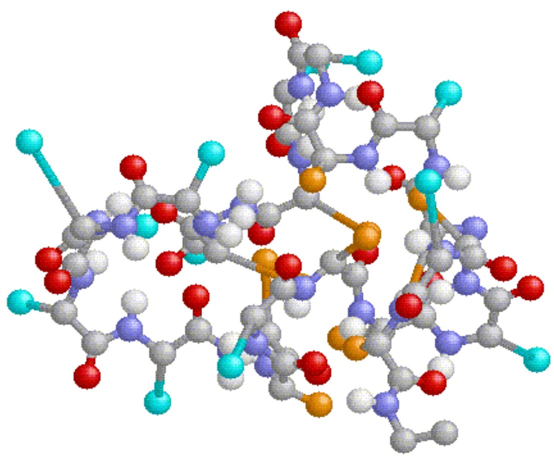
Time as a function of n for 16 PEs on the Cray T3D

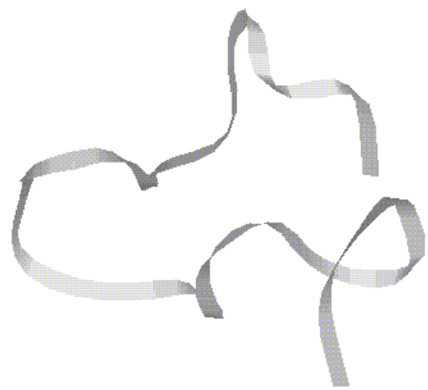
$T(n) \approx (0.01)\, n^4$

n	10	20	30	40	50	100
T(n) minutes	15	74	235	595	1293 (21 hrs)	17505 (12 days)

A.T. Phillips
J.B. Rosen
K.A. Dill

A 23-mer Folded Structure (BBA1 Motif)





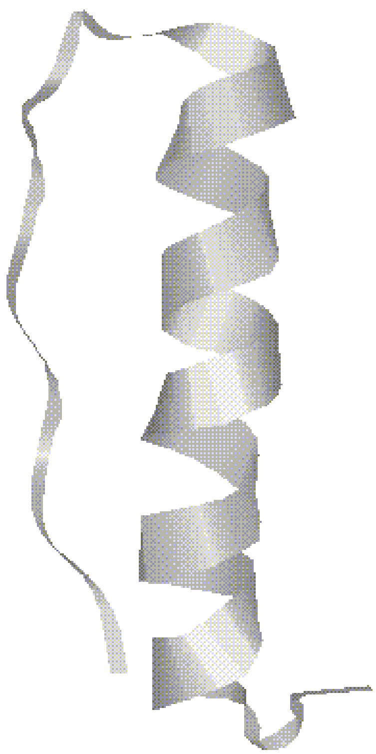
Computed Energy = -160.31 Kcal/mol

- Compare this structure to:

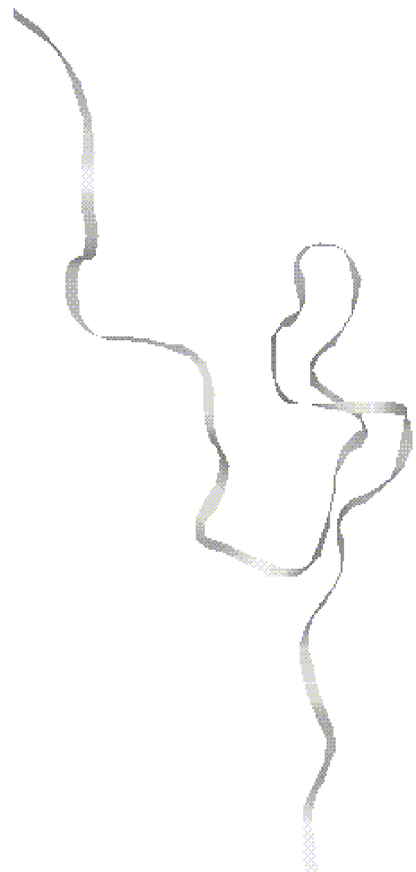
  M.D. Struthers, R.P. Cheng, and B. Imperiali, *Design of a Monomeric 23-Residue Polypeptide with Defined Tertiary Structure*, Science **271**:342-345 (19 January 1996)

A 36-mer Folded Structure (1PPT)

Native structure



CGU computed structure



Computed Energy = -309.94 Kcal/mol

Computed Energies of 8 Small Proteins

Compound Name	Residues	CGU Native Energy	Time for Solution[a]
met-enkephalin	5	-43.78 kcal/mol	1.2 minutes
bradykinin	9	-22.35 kcal/mol	6.5 minutes
oxytocin	9	-105.17 kcal/mol	3.3 minutes
BBA1	23	-160.31 kcal/mol	1.6 hours
mellitin	27	-262.69 kcal/mol	3.7 hours
zinc-finger motif	30	-153.06 kcal/mol	2.3 hours
avian pancreatic polypeptide	36	-306.94 kcal/mol	7.7 hours
crambin	46	-325.35 kcal/mol	8.0 hours[b]

a. Time reported is “wall clock time” using 16 nodes on the Cray T3D.

b. Time reported is “wall clock time “ using 32 nodes on the Cray T3D.

A.T. Phillips
J.B. Rosen
K.A. Dill

Probability of a Local Minimum Conformation

- Given $N+1$ local minima (including the global) with energies

$$F_j,\ j=0,\ldots,N$$

where $F_0 = F_G$ is the global minimum energy.

- The probability of the ith conformation is:

$$p_i = \frac{e^{-(F_i - F_G)/(kT)}}{\sum_{j=0}^{N} e^{-(F_j - F_G)/(kT)}}$$

where $k = 1.982$ cal/mol, and T = temperature (degrees Kelvin)

■
A.T. Phillips
J.B. Rosen
K.A. Dill

Probability Distribution of Local Minima

	Number of Local Minima in Probability Range Shown										
compound (residues)	.9	.8	.7	.6	.5	.4	.3	.2	.1	<.1	Total
met-enkephalin (5)					1			1		78	80
bradykinin (9)								1	2	116	119
oxytocin (9)			1				1			99	101
BBA1 (23)	1									302	303
mellitin (27)	1									383	384
zinc-finger motif (30)	1									320	321
avian pancreatic polypeptide (36)	1									609	610
crambin (46)	1									651	652

Interpretation of CGU Coefficients

Final "Landscape" CGU Energy Function:

$$\Psi(\phi) = F_G + \frac{1}{2}\sum_{i=1}^{n} d_i(\phi_i - (\phi_G)_i)^2$$

Holding all internal coordinates, except ϕ_j, fixed at $(\phi_0)_i$ gives:

$$\Psi(\phi_j) - F_G = \frac{1}{2} d_j[\phi_j - (\phi_G)_j]^2$$

The Boltzmann distribution gives the probability in terms of the energy:

$$P(\phi_j) = e^{-\frac{d_j}{2kT}[\phi_j - (\phi_G)_j]^2}$$

Therefore, $(\phi_G)_j$ is the mean value of ϕ_j, and $\sigma_j^2 = (kT)/d_j$ is its variance.

Representation of the Energy Landscape

The difference between the CGU energy and the global minimum energy is:

$$\Psi(\phi) - F_G = \frac{1}{2}\sum_{i=1}^{n} d_i[\phi_i - (\phi_G)_i]^2$$

The "RMS weighted error", the deviation of the φ_i from their global minimum values $(\varphi_G)_i$, is:

$$\overline{\Delta\phi} = \sqrt{\sum_{i=1}^{n} d_i[\phi_i - (\phi_G)_i]^2}$$

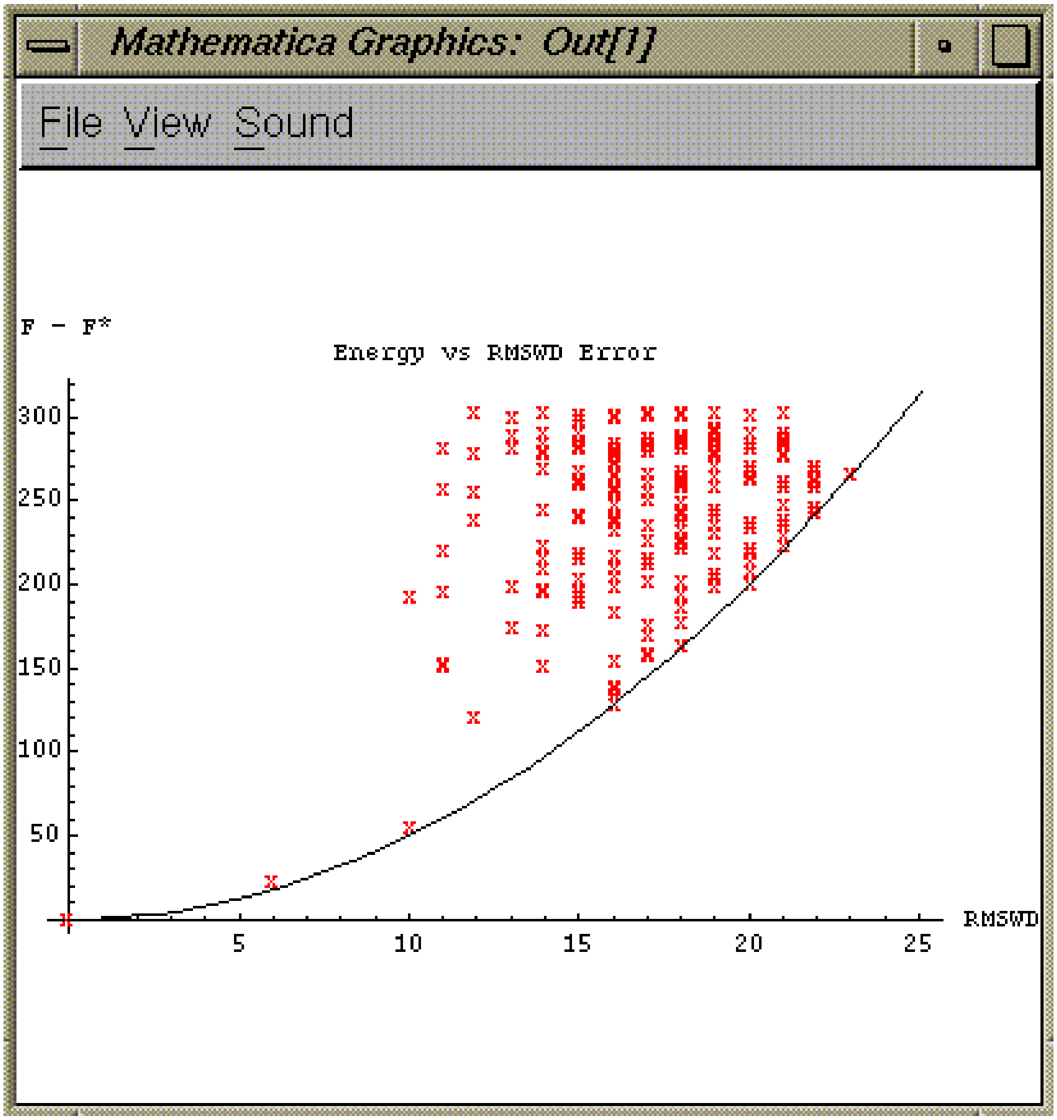
Hence: $\Psi(\phi) - F_G = \frac{1}{2}(\overline{\Delta\phi})^2$.

Plotting $F(\phi)$-F_G vs $\overline{\Delta\phi}$ gives a representation of the energy landscape.

A.T. Phillips
J.B. Rosen
K.A. Dill

Energy Landscape (1PPT)

Distribution of Local Minima



A.T. Phillips
J.B. Rosen
K.A. Dill

Effect of Sequence on Structure

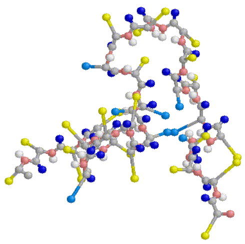
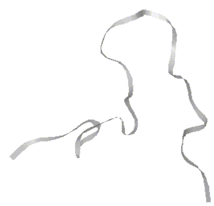
- The primary sequence uniquely determines the folded structure.
- Permutations of the primary sequence result in dramatically different structures.
- Permutations of the sequence do not significantly affect the computational efficiency of the CGU method.

Five Permutations of a 30-mer Sequence (27% Hydrophobic)

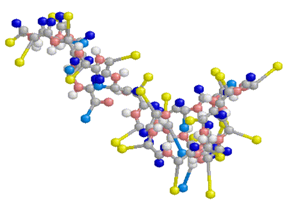
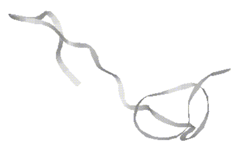
Sequence	Wall Time	Passes	Time/Pass	Min Energy
Seq1	224 m	3	75 m	-118.14
Seq2	323 m	6	81 m	-127.26
Seq3	208 m	3	69 m	-107.71
Seq4	139 m	9	70 m	-90.64
Seq5	332 m	5	83 m	-157.96
Avg	245 m	5.2	75 m	-120.34

A.T. Phillips
J.B. Rosen
K.A. Dill

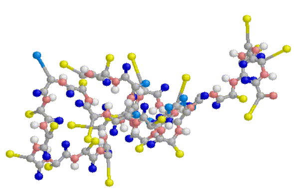
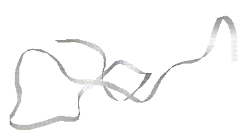
Effect of Sequence on Structure (cont)

$\Psi(\phi) = -118.14$ kcal/mol

$\Psi(\phi) = -107.71$ kcal/mol

$\Psi(\phi) = -157.96$ kcal/mol

■
A.T. Phillips
J.B. Rosen
K.A. Dill

Relationship to Folding Dynamics

- The CGU can be represented as:

$$\Psi(\phi) - F_G = \frac{1}{2}\sum_{i=1}^{n} d_i[\phi_i - (\phi_G)_i]^2.$$

- Starting with any initial conformation $\phi^{(0)}$, we assume that the ϕ_i, as a function of time t, are determined by the steepest descent path on Ψ.

- This is given by the ODE system:

$$\frac{d\phi}{dt} = -\mu\nabla\Psi(\phi),\ t \geq 0,\ \phi(0) = \phi^{(0)}$$

where μ is a rate constant.

- Combining these two equations gives:

$$\frac{d\phi_i}{dt} = -\mu d_i[\phi_i - (\phi_0)_i],\ \phi_i(0) = (\phi^{(0)})_i,\ i=1,\ldots,n.$$

A.T. Phillips
J.B. Rosen
K.A. Dill

Relationship to Folding Dynamics (cont)

- This has the obvious solution:

$$\phi_i(t) - (\phi_G)_i = \left[(\phi^{(0)})_i - (\phi_G)_i \right] e^{-\mu d_i t}, \; t \geq 0, \; \text{i=1,...,n}.$$

- Hence, as t increases, each ϕ_i will approach $(\phi_\mathrm{G})_i$ at a rate determined by μd_i.

- And the potential energy can then be expressed as:

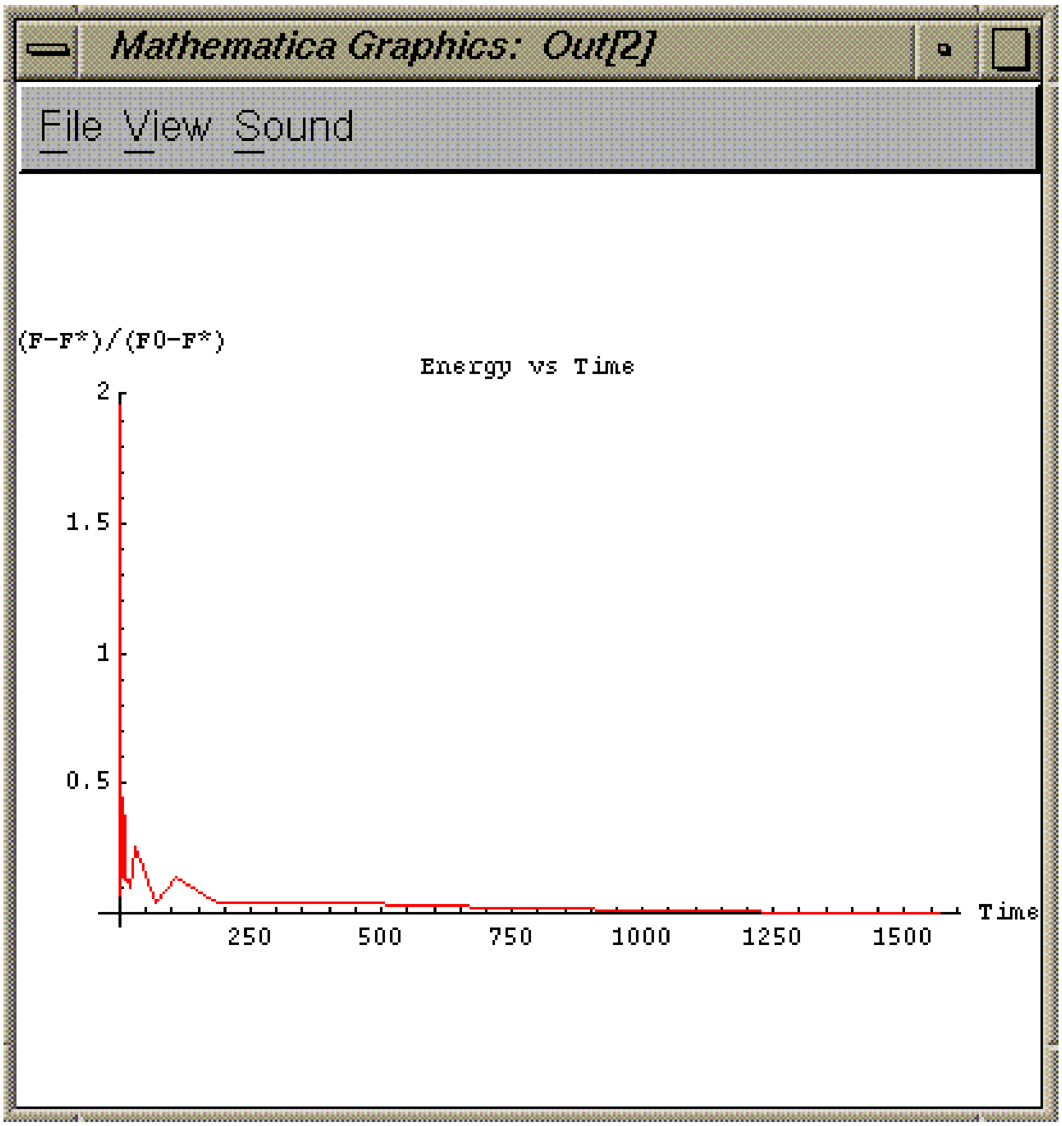
$$\Psi(\phi(t)) = \frac{1}{2} \sum_{i=1}^{n} d_i \left[(\phi^{(0)})_i - (\phi_G)_i \right]^2 e^{-2\mu d_i t} + F_G.$$

- The CGU surface $\Psi(\phi)$ is a smoothed approximation to the "energy funnel" which determines the folding dynamics.

A.T. Phillips
J.B. Rosen
K.A. Dill

Example Folding Dynamics (1PPT)

Potential Energy Plot



Coordinate Translation

- The computed global solution ϕ_G may not coincide with the known native structure ϕ_N.
- A simple coordinate translation can be used to map the computed global minimum structure to the known native structure.
- Define $\Delta\phi_N = \phi_G - \phi_N$ and the translated energy function:

$$\bar{F}(\phi) = F(\phi + \Delta\phi_N).$$

- Note: $\bar{F}(\phi_N) = F(\phi_G)$ so that $\bar{F}(\phi)$ has its global minimum at ϕ_N.

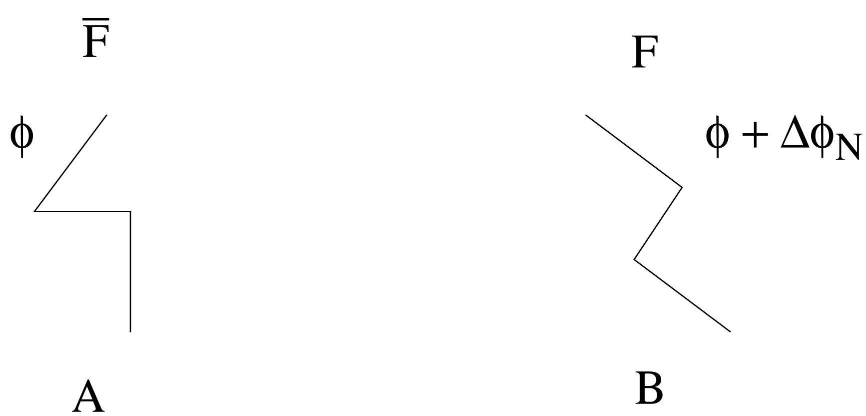
■
A.T. Phillips
J.B. Rosen
K.A. Dill

Coordinate Translation (cont)

- Also:

$$\bar{F}(\phi^{(j)} - \Delta\phi_N) = F(\phi^{(j)}), \text{ for j=1,...,k.}$$

- Thus, $\bar{F}(\phi)$ will have a local minimum at each conformation $\phi^{(j)} - \Delta\phi_N$, j=1,...,k (these are the translated local minima).
- The energy $\bar{F}(\phi)$ is given by the original energy *F* at a *different conformation* $\phi - \Delta\phi_N$.



A.T. Phillips
J.B. Rosen
K.A. Dill