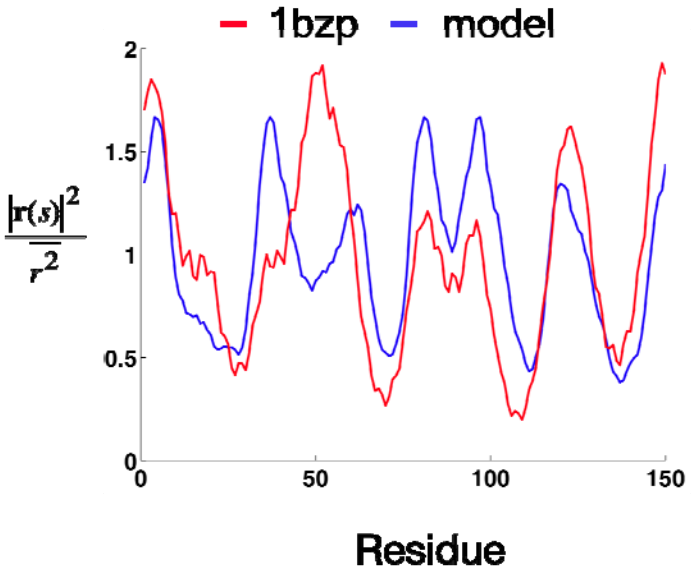
Structure *19*

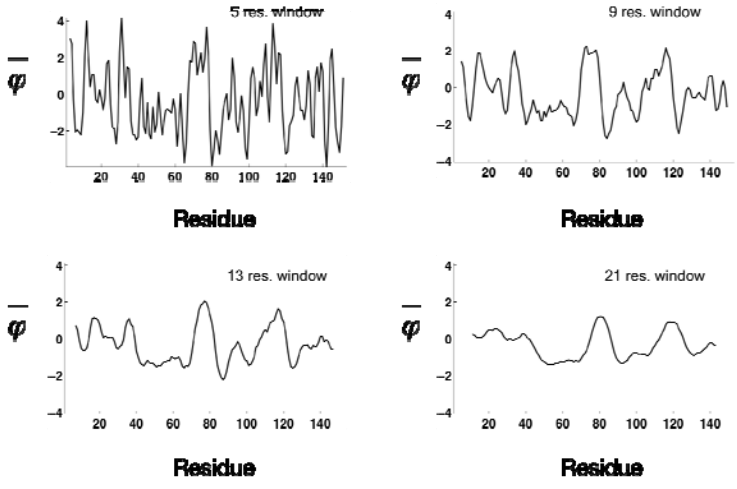
Supplemental Information

Allostery in Protein Domains Reflects a Balance of Steric and Hydrophobic Effects

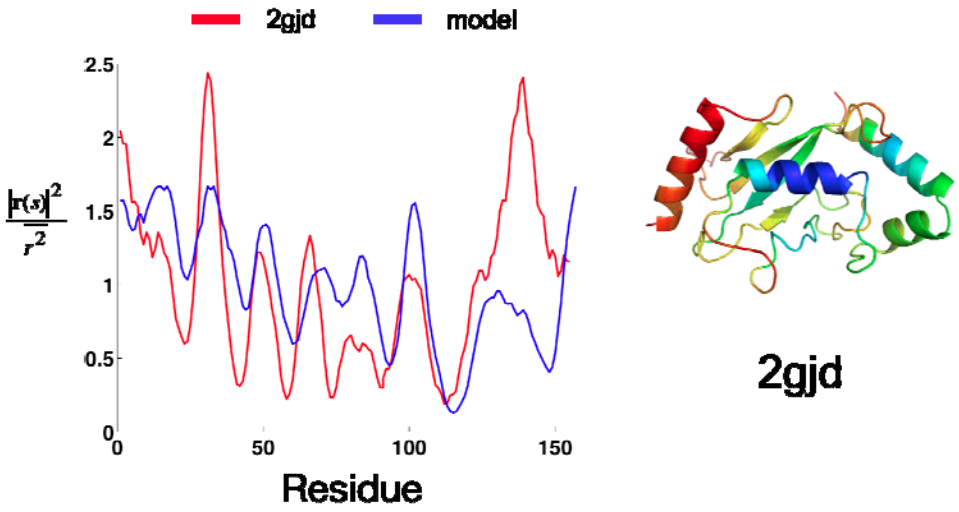
Jeremy L. England

a





b


2gjd
model
$\frac{|r(s)|^2}{r^2}$
2.5
2
1.5
1
0.5
0
0
50
100
150
Residue
2gjd

c

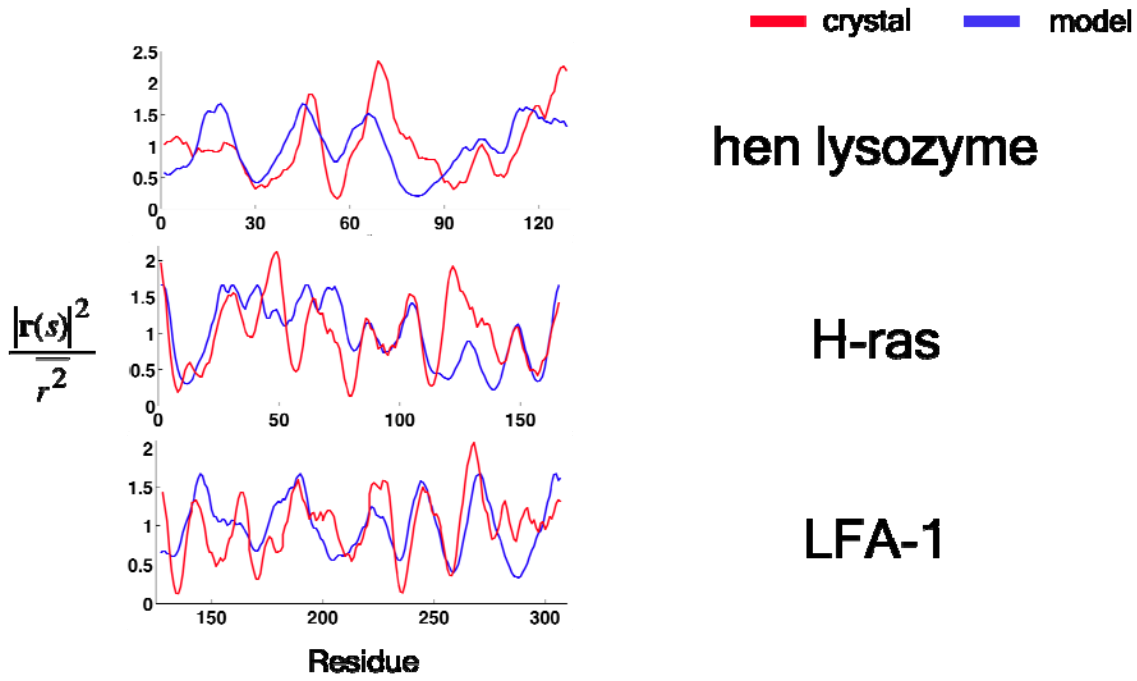
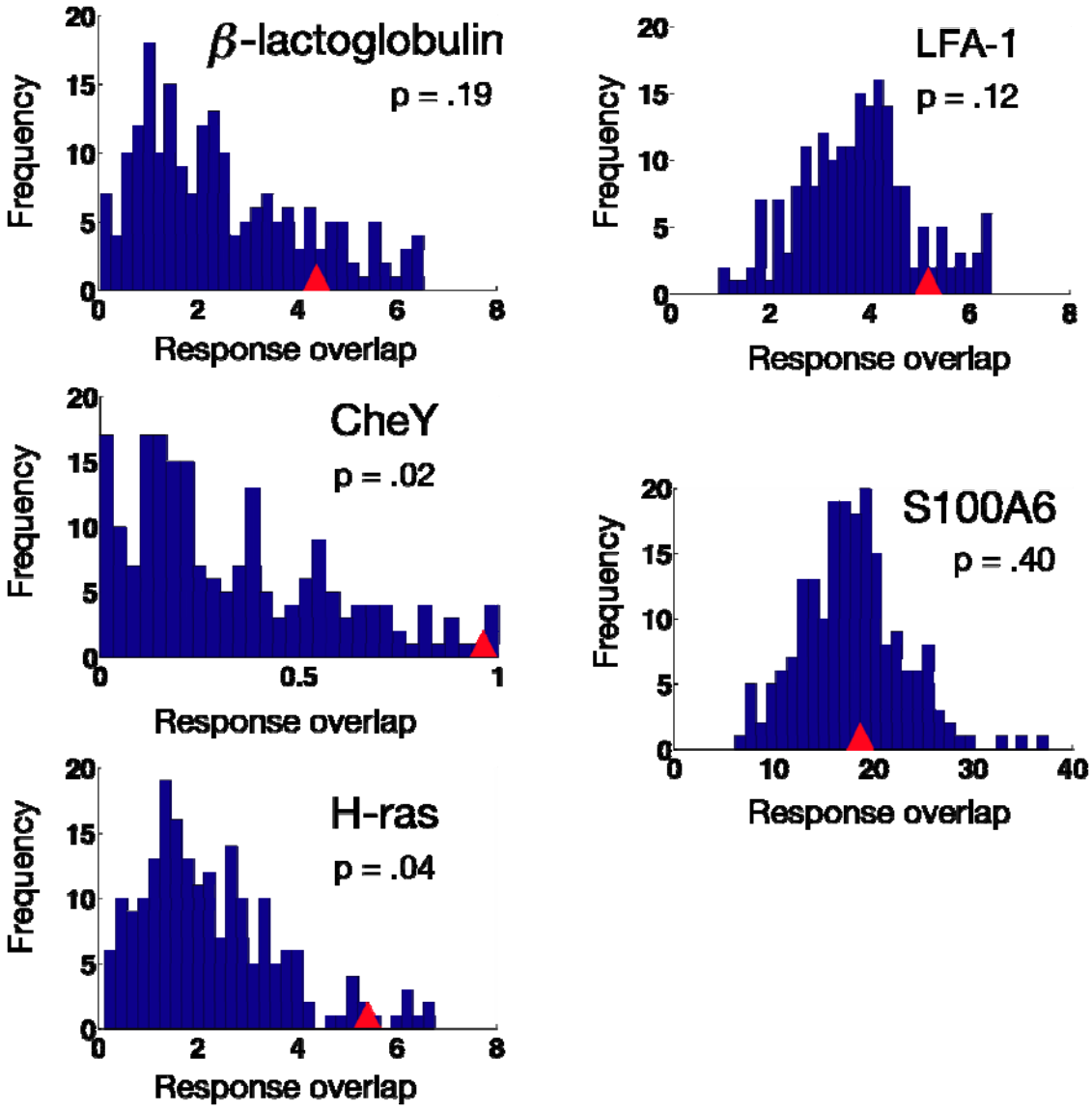

crystal
model
$\frac{|r(s)|^2}{r^2}$
2.5
2
1.5
1
0.5
0
0
30
60
90
120
0
50
100
150
150
200
250
300
Residue
hen lysozyme
H-ras
LFA-1

Figure S1. a The myoglobin burial traces (top panel) calculated from the model (blue) and computed from the crystal structure PDB ID 1BZP (red) in Figure 1 are compared to the window-averaged Kyte-Doolittle hydropathy of the myoglobin sequence (bottom panel) for four different window sizes. **b** The UBC9 burial traces (top panel) calculated from the model (blue) and computed from the crystal structure (red) are plotted, and the model trace is used to color the PDB ID 2GJD crystal structure (blue most buried, red the least buried) in the right hand side. **c** The burial traces calculated from the model (blue) and computed from the crystal structure (red) are plotted for hen lysozyme, H-ras, and LFA-1.

a


β-lactoglobulin
p = .19
LFA-1
p = .12
CheY
p = .02
S100A6
p = .40
H-ras
p = .04
Frequency
Response overlap

b

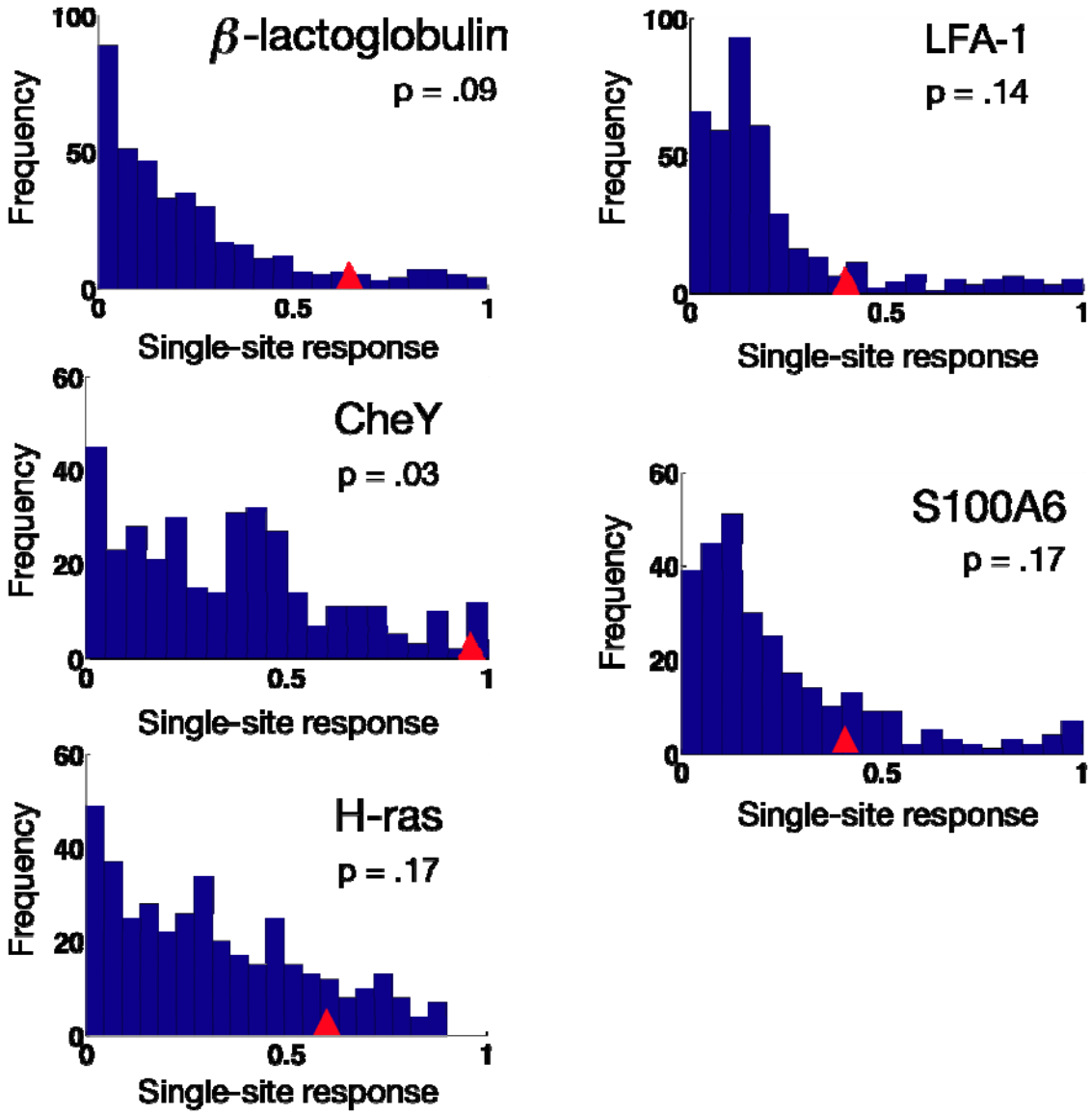


Figure S2 **a** Control distributions for allosteric response were generated by scanning random permutations of the starting sequence for sequences whose predicted burial trace had a correlation of 0.4 or higher with the target structure. These sequences were then used for each protein to generate a burial covariance matrix from 20 randomly sampled configurations at 1 kT above the minimum, and the resulting response was generated for the corresponding inputs (green triangles in Figure 3), and normalized to the maximum height it achieved outside of the region within 3 residues of an input residue along the chain. For each protein, a response overlap was then computed from the sum of the normalized response over all output residues in the protein (red squares in Figure 3). Here the distributions of response overlap from 200 samples are plotted for the proteins in question, with the red triangle indicating the response overlap computed for the true sequence. The p-value reported is the fraction of sequences that fall to the right of the red arrow in each distribution. **b** Distributions were also generated to test the strength

of the measured allosteric response on the wild-type sequences. Each of the five sequences in Figure 3 was used to generate a burial covariance matrix at 1 kT above the minimum, and then pairs of residues on the chain more than 20 residues apart were randomly selected (with 500 draws, one site being designated as "stimulus," and the other "response") to compute the absolute magnitude of their burial covariance normalized to the maximum strength of covariance along the chain not within three residues of the stimulus site. These distributions of single-site responses were compared to the average response per residue (red arrows) calculated for each of the allosteric systems in Figure 3. The p-value reported is the fraction of pair samples that fall to the right of the red arrow in each distribution.

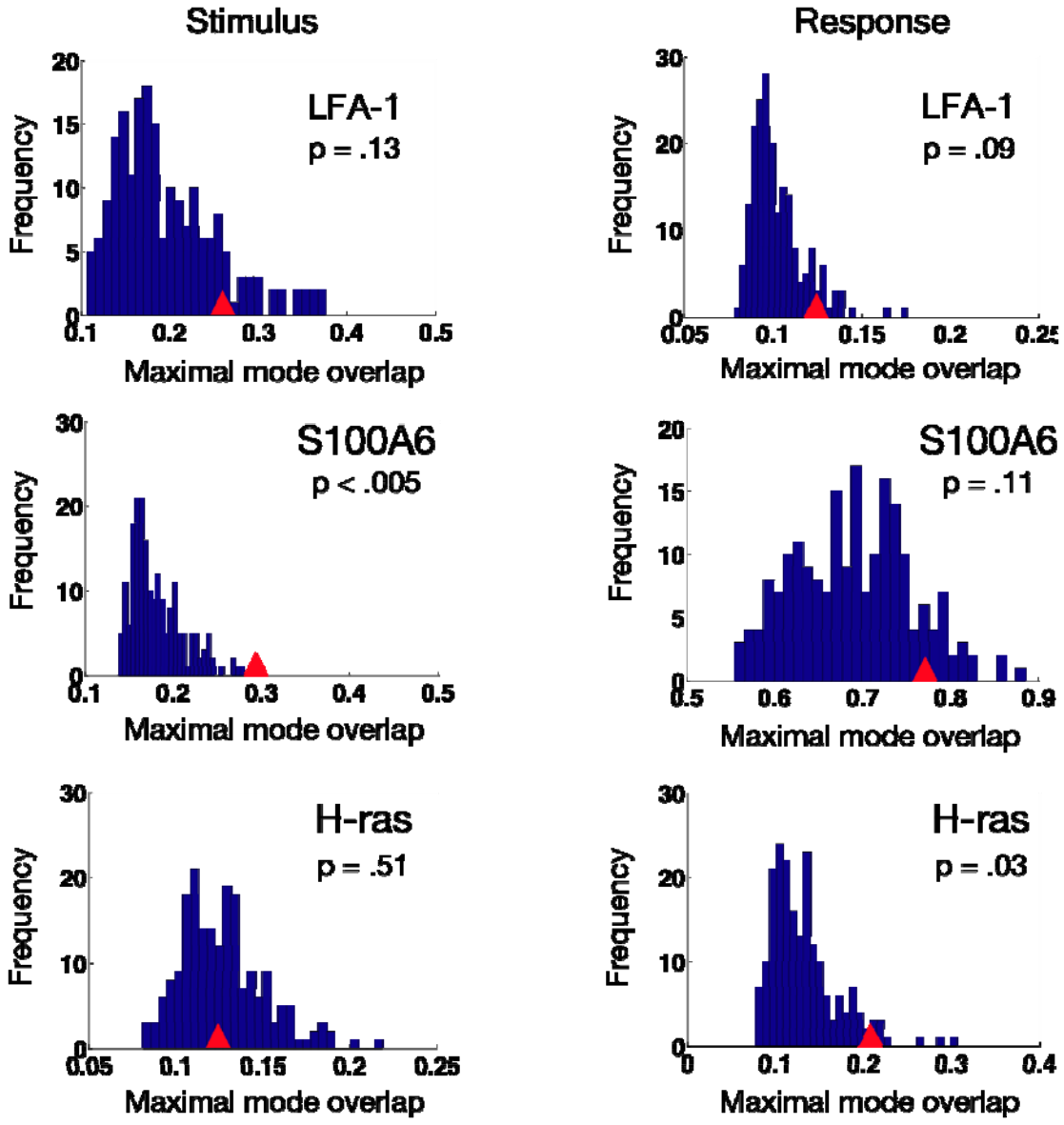


Figure S3 Control distributions for the overlap of allosteric sites with specific sequence burial modes were generated by scanning random permutations of the starting sequence for sequences whose predicted burial trace had a correlation of 0.4 or higher with the target structure. These sequences were then used for each

protein to compute burial modes, and by integrating each mode over the stimulus sites (green triangles in Figure 4) and response sites (red squares in Figure 4), a distribution was generated of the overlap between allosteric site and the burial mode that maximally overlapped for each sequence. The red triangle in each panel indicates the overlap of the allosteric site in question with the corresponding mode identified in Figure 4. The p-value reported is the fraction of sequences that fall to the right of the red arrow in each distribution.

Supplemental Experimental Procedures

Burial Trace Approximation

The burial trace in the model is exactly given in terms of the coefficients of the burial eigenbasis by

$$\left|\mathbf{r}(s)\right|^2 = \sum_k c_k \psi_k(s)^2 + \sum_{j \neq k} \left(X_j X_k + Y_j Y_k + Z_j Z_k \right) \psi_j(s) \psi_k(s)$$

At this point, it is necessary to make a crucial approximation. It is clear that a given choice of the constants c_k specifies a sub-ensemble of conformations that vary based on the specific values of the cross-terms in the sum above. All such conformations will have the same energy and the same mean-square radius, and, if energy is to any degree determined by particular features of tertiary structure, it is reasonable to suppose that conformations within a sub-ensemble are similar enough to each other that the average value of the distance trace $\left|\mathbf{r}(s)\right|^2$ within the sub-ensemble can be taken to be representative of the sub-ensemble as a whole. This is especially the case since the cross-terms must integrate to zero over s and therefore make no average contribution to $\left|\mathbf{r}(s)\right|^2$. Following this line of reasoning leads to the ansatz:

$$R^2 \geq \left|\mathbf{r}(s)\right|^2 \approx \sum_k c_k \psi_k(s)^2$$

MATLAB Code used for linear programming energy optimization:

```
function [ copt,V,eps,emin,r2max,msr,phi0,seq] = pholder(FASTA)

%Start with an amino acid sequence string denoted by the variable FASTA.
%The code computes the optimal square-distance of backbone from the center
%of the protein, and plots it, returning the optimal weights for the
%burial modes, and the matrix of modes.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
k = 1.5;            % This defines the spring stiffness between monomers in
                    % the chain. It is chosen so that a free Gaussian coil
                    % has mean square inter-monomer distance of 1, so that
                    % other distances are measured in units of the distance
                    % between two adjacent alpha carbons in the backbone.

M = 10000000;       % This weight simply ensures that the center of mass of
                    % the globule remains fixed at the origin.

n = length(FASTA);    % The number of residues in the amino acid sequence.

rho0 = 250/(4*pi*(4^3)/3);  % Estimate of monomer density in proteins
                            % using TIM barrel as a benchmark

r2max = (3*n/rho0/4/pi)^(2/3);% Estimate of squared max radius of globule.

phi0 = (250/n)^(2/3)/112.5;      % The energy scale of hydropathy.
                                  % Here, the change in globule size with
                                  % sequence length affects the magnitude
                                  % of the hydropathy scale used in the
                                  % theory since the transfer energy from
                                  % surface to core should be roughly the same
                                  % regardless of globule size (rmax).
%%%%%%%%text%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Now construct the matrix of harmonic bonds for the polymer.

bond = 0;

for i = 1:n,
    for j = 1:n,
        bond(i,j) =  (-((j + 1) == i) - ((i + 1) == j) + 2* (i == j))*k+M;
    end
end


% Leaving the ends of the polymer free to flop around


bond(1,1) = bond(1,1) - k;
bond(n,n) = bond(n,n) - k;

% Now turn the sequence into standard Kyte-Doolittle hydropathies

seq = 0;

for i = 1:n,
    seq(i) = KD(FASTA(i));
end

% Now construct the Hamiltonian H and diagonalize it

H = bond + phi0 * diag(seq);

[psi, epsmat] = eig(H);    % Psi are the eigenvectors, eps the eigenvalues

eps = diag(epsmat);

% Now construct the matrix V of squared amplitudes for the eigenfunctions

V = psi.^2;

% Now compute the optimal backbone trace using linear programming.
```

```
lb = 0*ones(n,1); % lower bound vector preventing
                  % c values from being negative

msr = 3*r2max/5;  % mean square radius fixed in ratio to r2max

[copt,emin] = linprog(eps,V,ones(n,1)*r2max,ones(1,n),msr * n,lb,[],[]);


optstruc = V*copt;

end
```