Theor Chem Acc (1999) 103:77–80
DOI 10.1007/s00214990m192

Theoretical
Chemistry Accounts

Letter

Using simplified protein representation as a reference potential for all-atom calculations of folding free energy

Z.Z. Fan[1], **J.-K. Hwang**[1], **A. Warshel**[2]

[1] Department of Life Sciences, National Tsing Hua University, Hsin Chu, Taiwan
[2] Department of Chemistry, University of Southern California, Los Angeles, CA 90089, USA

Received: 30 December 1998 / Accepted: 27 January 1999 / Published online: 5 May 1999

Abstract. An effective approach for evaluating folding free-energy surfaces of explicit all-atom models is developed and examined. This approach is based on using the potential of a simplified protein model as a reference potential for calculating the free energy of the corresponding explicit model. Preliminary results are presented for the folding free energy of a 12-residue helix. The potential of the method for studies of protein-folding processes is discussed, emphasizing the ability to determine the difference between the results of simplified and explicit models. This can help in establishing the validity of simplified folding models.

Key words: Protein folding – Simplified models

1 Introduction

Many current simulations of protein folding involve the use of a simplified representation where each sidechain is replaced by a single interaction center [1–6]. While this representation is extremely effective, one would like to know what features of the folding process are missing in such a seemingly oversimplified approach. One way to address this question is of course to try to use an explicit all-atom model [7–9], but this approach is extremely expensive and is not yet practical for all but small proteins.

In this work we develop an approach that allows one to explore the question of the relationship between the simplified and the all-atom model and also provide an effective way of evaluating the folding free-energy surface of all-atom models. Our approach is based on the use of a simplified model as a reference potential and then using the difference between the all-atom and simplified potentials to obtain the all-atom free energy. This approach is similar in some respects to the classical quantized path approach used by us in studying quantum mechanical nuclear effects in chemical reactions in condensed phases [10, 11]. The present paper formulates our approach and presents preliminary results that explore its effectiveness. This is done by evaluating the free-energy curve for the folding of a 12-residue helix.

2 Methods

As seen from Fig. 1, a protein molecule can be represented by an explicit model and a coordinate set **r** or a simplified representation and a coordinate set **R**. The coordinate set **R** and a set of rigid rotation coordinates Φ can be used to generate the coordinate set **r**. The potential surfaces (force fields) that correspond to the simplified and the explicit representations will be referred to here as U_{sp} and V_{ep}, respectively. Now it is quite practical to use U_{sp} to estimate the free-energy surfaces for the folding process. This can be done by applying a free-energy perturbation coupled with an umbrella sampling approach. The corresponding free energy will be expressed as [12]

$$\exp\left[-\beta\Delta g_{\mathrm{sp}}(X)\right] = \exp[-\beta\Delta G(\lambda_m)]\langle\delta(X'-X) \times \exp\{-\beta[U_{\mathrm{sp}}(X') - U_m(X')]\}\rangle_m \tag{1}$$

$$\exp[-\beta\Delta G(\lambda_m)] = \sum_{m'=0}^{m-1} \ln\langle\exp\{-\beta[U_{m'+1} - U_{m'}]\}\rangle_{m'} \ ,$$

where X is the folding reaction coordinate, β is $1/k_{\mathrm{B}}T$ with k_{B} and T denoting the Boltzmann constant and temperature, respectively. The delta function $\delta(X'-X)$ is assigned a value of 1 when $X' = X$, and is 0 otherwise. The average $\langle\ldots\rangle_m$ designates an average over the U_m mapping potential, which is given by

$$U_m = U_{\mathrm{sp}} + V_{\mathrm{c}}(\lambda_m) \ , \tag{2}$$

In the present work the constraint potential $V_{\mathrm{c}}(\lambda_m)$ is given by

$$V_{\mathrm{c}} = (1-\lambda_m)K_{\mathrm{c}}\left(R_{\mathrm{g}} - R_{\mathrm{g}}^{\mathrm{i}}\right)^2 + \lambda_m K_{\mathrm{c}}\left(R_{\mathrm{g}} - R_{\mathrm{g}}^{\mathrm{f}}\right)^2 \ , \tag{3}$$

where λ_m is the mapping parameter, K_{c} the constraint force constant, R_{g} the radius of gyration of the protein, and $R_{\mathrm{g}}^{\mathrm{i}}$ and $R_{\mathrm{g}}^{\mathrm{f}}$ are the initial and final constraint radii of gyration, respectively. The initial constraint radius of gyration, $R_{\mathrm{g}}^{\mathrm{i}}$, is set to that of a native protein and the final constraint radius of gyration, $R_{\mathrm{g}}^{\mathrm{f}}$, is set to a larger value to induce unfolding. During simulation, the value of λ_m

Correspondence to: A. Warshel

changes discretely from 0 to 1, thus driving the protein from the native state to the unfolded state. In this work, we choose the radius of gyration as the reaction coordinate.

Trying to apply the same approach used in Eq. (1) to obtain the free-energy function for the folding process in the all-atom representation is very demanding and might encounter major convergence problems. Here, however, one may exploit a trick used in several related problems such as calculations of quantum mechanical free energies of reactions in condensed phases [13–15] and in evaluating nuclear quantum mechanical effects in condensed phases [10, 11, 16]; that is, we consider the cycle of Fig. 2 where the all-atom free-energy function $\Delta g_{\mathrm{ep}}(X)$ is obtained by using $\Delta g_{\mathrm{sp}}(X)$ and then we evaluate the free-energy changes of moving from U_{sp} to V_{ep} in selected points along the reaction coordinate.

The free-energy difference $\Delta\Delta g_{\mathrm{sp}\to\mathrm{ep}}$ is obtained by

$$\Delta\Delta g_{\mathrm{sp}\to\mathrm{ep}}(X) = \frac{q_{\mathrm{ep}}(X)}{q_{\mathrm{sp}}(X)} \ , \tag{4}$$

where $q_{\mathrm{ep}}(X)$ and $q_{\mathrm{sp}}(X)$ are the contributions to the corresponding total partition functions Q_{ep} and Q_{sp} from the given X so that $Q = \int q(X)\,\mathrm{d}X$. For convenience, we will first consider the ratio between the total partition functions, since the same derivations can also be applied to Eq. (4).

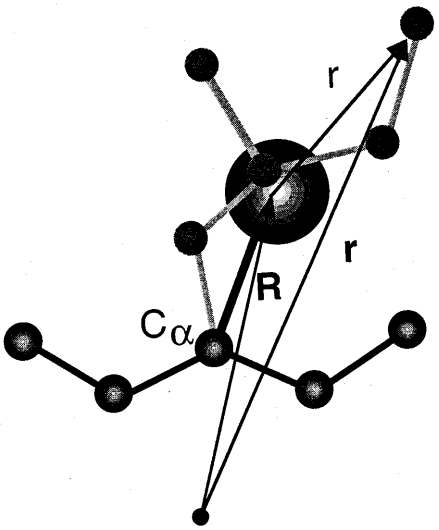

Fig. 1. A schematic illustration of the relation between the simplified and the explicit models. **R** and **r** indicate the coordinates of the simplified and the explicit models, respectively

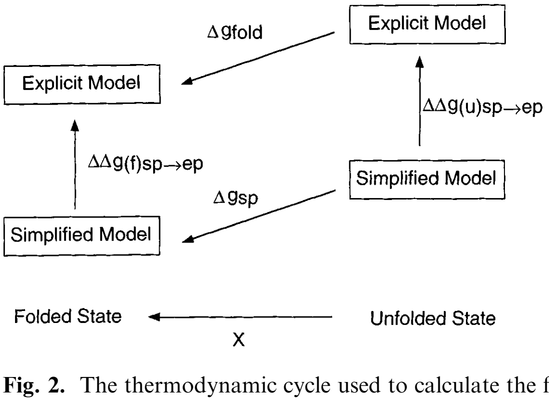

Fig. 2. The thermodynamic cycle used to calculate the folding free energy Δg_{fold}. The free energy change Δg_{sp} from the unfolded to the folded simple model is calculated by the free-energy perturbation method. The free-energy changes $\Delta\Delta g_{(\mathrm{f})\mathrm{sp}\to\mathrm{ep}}$ and $\Delta\Delta g_{(\mathrm{u})\mathrm{sp}\to\mathrm{ep}}$ are obtained by umbrella sampling over X, the reaction coordinate

$$\frac{Q_{\mathrm{ep}}}{Q_{\mathrm{sp}}} = \frac{\int \mathrm{d}\mathbf{R} \int \mathrm{d}\mathbf{r} \exp[-\beta V_{\mathrm{ep}}(\mathbf{R},\mathbf{r})]}{\int \mathrm{d}\mathbf{R} \int \mathrm{d}\mathbf{r} \exp[-\beta U_{\mathrm{sp}}(\mathbf{R})]} \ , \tag{5}$$

where V_{sp}, the potential for the explicit model, and U_{sp}, the potential for the simplified models, are, respectively,

$$V_{\mathrm{ep}}(\mathbf{R},\mathbf{r}) = V'_{\mathrm{mc-mc}}(\mathbf{R}) + V_{\mathrm{sc}}(\mathbf{r};\mathbf{R}) \tag{6a}$$

$$U_{\mathrm{sp}}(\mathbf{R}) = V'_{\mathrm{mc-mc}}(\mathbf{R}) + U_{\mathrm{sc}}(\mathbf{R}) \ , \tag{6b}$$

where $V_{\mathrm{sc}}(\mathbf{r};\mathbf{R})$ and $U_{\mathrm{sc}}(\mathbf{R})$ are the other interactions such as main chain–sidechain, sidechain–sidechain interactions and others for the explicit and the simplified models, respectively.

$$\begin{aligned}\frac{Q_{\mathrm{ep}}}{Q_{\mathrm{sp}}} &= \frac{\int \mathrm{d}\mathbf{R}\,\mathrm{d}\mathbf{r} \exp\{-\beta[V'_{\mathrm{mc-mc}}(\mathbf{R}) + V_{\mathrm{sc}}(\mathbf{r};\mathbf{R})]\}}{\int \mathrm{d}\mathbf{R}\,\mathrm{d}\mathbf{r} \exp\{-\beta[V'_{\mathrm{mc-mc}}(\mathbf{R}) + U_{\mathrm{sc}}(\mathbf{R})]\}} \\ &= \frac{\int \mathrm{d}\mathbf{R} \exp(-\beta U_{\mathrm{sp}}) \mathrm{d}\mathbf{r} \exp[-\beta(V_{\mathrm{sc}} - U_{\mathrm{sc}})]}{\int \mathrm{d}\mathbf{R} \exp(-\beta U_{\mathrm{sp}}) \int \mathrm{d}\mathbf{r}} \\ &= \left\langle \int \mathrm{d}\mathbf{r} \exp[-\beta(V_{\mathrm{sc}} - U_{\mathrm{sc}})] \Big/ \int \mathrm{d}r \right\rangle_{\mathrm{sp}} \\ &= \left\langle \langle \exp[-\beta(V_{\mathrm{sc}} - U_{\mathrm{sc}})] \rangle_{\mathrm{fp}} \right\rangle_{\mathrm{sp}}\end{aligned} \tag{7}$$

The notation $\langle \ldots \rangle_{\mathrm{fp}}$ represents an average over the coordinates of explicit atoms in each sidechain with a constraint potential $V_{\mathrm{fp}} = 0$ (note that fp designates free potential). The term $\langle \exp[-\beta(V_{\mathrm{sc}} - U_{\mathrm{sc}})] \rangle_{\mathrm{fp}}$ can also be written as

$$\begin{aligned}\langle \exp[-\beta(V_{\mathrm{sc}} - U_{\mathrm{sc}})] \rangle_{\mathrm{fp}} &= \frac{\int \mathrm{d}\mathbf{r} \exp\{-\beta[V_{\mathrm{sc}}(\mathbf{r};\mathbf{R}) - U_{\mathrm{sc}}(\mathbf{R})]\}}{\int \mathrm{d}r} \\ &= F \exp[\beta U_{\mathrm{sc}}(\mathbf{R})] \ ,\end{aligned} \tag{8}$$

where

$$\begin{aligned}F &= \frac{\int \mathrm{d}\mathbf{r} \exp[-\beta V_{\mathrm{sc}}(\mathbf{r},\mathbf{R})]}{\int \mathrm{d}r} \\ &= \frac{\int \mathrm{d}\mathbf{r} \exp[-\beta V_{\mathrm{sc}}(\mathbf{r},\mathbf{R}) - 0]}{\int \mathrm{d}r \exp[-\beta 0]} = \langle \exp(-\beta V_{\mathrm{sc}}) \rangle_{\mathrm{fp}} \ .\end{aligned} \tag{9}$$

Finally we obtain

$$\frac{Q_{\mathrm{ep}}}{Q_{\mathrm{sp}}} = \left\langle \langle \exp[-\beta(V_{\mathrm{sc}} - U_{\mathrm{sc}})] \rangle_{\mathrm{fp}} \right\rangle_{\mathrm{sp}} = \langle F \exp[\beta U_{\mathrm{sc}}] \rangle_{\mathrm{sp}} \ . \tag{10}$$

Returning to Eq. (2), we have

$$\begin{aligned}\exp[-\beta \Delta g_{\mathrm{ep}}(X)] &= \frac{q_{\mathrm{ep}}(X)}{Q_{\mathrm{ep}}} = \frac{Q_m}{Q_{\mathrm{ep}}} \frac{q_{\mathrm{ep}}(X)}{Q_m} \\ &= \left\langle \langle \exp[-\beta(V_{\mathrm{ep}} - U_m)] \rangle_{\mathrm{fp}} \right\rangle_m \\ &\quad \times \left\langle \langle \delta(X - X') \exp\{-\beta[V_{\mathrm{ep}}(X') - U_m(X')]\} \rangle_{\mathrm{fp}} \right\rangle_m\end{aligned} \tag{11}$$

Similarly we have

$$\begin{aligned}\exp[-\beta \Delta g_{\mathrm{sp}}(X)] &= \left\langle \langle \exp[-\beta(V_{\mathrm{sp}} - U_m)] \rangle_{\mathrm{fp}} \right\rangle_m \\ &\quad \times \left\langle \langle \delta(X - X') \exp\{-\beta[U_{\mathrm{sp}}(X') - U_m(X')]\} \rangle_{\mathrm{fp}} \right\rangle_m \ .\end{aligned} \tag{12}$$

We can also write

$$\Delta\Delta g_{\mathrm{sp}\to\mathrm{ep}} = \Delta g_{\mathrm{ep}}(X) - \Delta g_{\mathrm{sp}}(X) \tag{13}$$

$$\begin{aligned}&\Delta\Delta g_{\mathrm{sp}\to\mathrm{ep}}(X) \\ &\quad = \left\langle \langle \delta(X - X') \exp[-\beta(V_{\mathrm{ep}}(X') - U_{\mathrm{sp}}(X'))] \rangle_{\mathrm{fp}} \right\rangle_{\mathrm{sp}} \ .\end{aligned} \tag{14}$$

Since in our simulation, the trajectories are propagated on U_m instead of on U_{sp}, our final equation is

$$\begin{aligned}&\Delta\Delta g_{\mathrm{sp}\to\mathrm{ep}}(X) \\ &\quad = \frac{\left\langle \langle \delta(X - X') \exp[-\beta(V_{\mathrm{ep}}(X') - U_{\mathrm{sp}}(X'))] \rangle_{\mathrm{fp}} \right\rangle_m}{\left\langle \langle \delta(X - X') \exp[-\beta(U_{\mathrm{sp}}(X') - U_m(X'))] \rangle_{\mathrm{fp}} \right\rangle_m} \ .\end{aligned} \tag{15}$$

For convenience, we used the radius of gyration as the reaction coordinate X in our simulation, but it should be noted that other choices of reaction coordinates can also be used in our approach. The simplified protein model originally introduced by Levitt and Warshel [1, 17] was used with some modifications in our simulations. In this model, the sidechain groups are approximated by effective van der Waals spheres while the atoms of the backbone are treated explicitly. The sidechain atoms interact with each other through an 8–6 interaction [1, 17], The original nonbonded parameters were further refined by the BFGS optimization [18–21]. The parameters were refined by minimizing the root-mean-square deviations between the calculated and observed values of both the atom positions and the protein sizes, i.e., the radii of gyration. The optimized nonbonded parameters are given in Table 1. The free-potential (fp) mapping of the all-atom representation was done by varying the torsional coordinates of the sidechains, while keeping the bond lengths and bond angles their standard values. The force field of the explicit all-atom model involves the standard ENZYMIX parameters [22] except that the nonpolar hydrogens were considered in the united atom representation. The effect of the solvent was considered implicitly by using the noniterative Langevin dipoles solvent model [23]. This model also includes a field-dependent hydrophobic term. The Langevin dipole solvent model was used for both the explicit and simplified protein models. The addition of the Langevin dipole model to the simplified-potential model is expected to change the folding energy but to have a smaller effect on the folded structure. Thus we kept the parameter sets that were obtained without the Langevin dipole model and these were refined using only structural information. A more consistent parametrization procedure that will include the simplified-potential and Langevin dipole model is left for subsequent studies.

3 Results

To examine the effectiveness of our approach we took as a test case the folding of a 12-mer helix, Gly-Trp-Glu-Ile-Pro-Glu-Pro-Tyr-Val-Trp-Asp-Glu. As a first stage we performed simulations using the simplified model. The simulations involved trajectories of 1fs-time steps at 300 K and a total time of 2 ns. The free-energy curve $\Delta g_{\mathrm{sp}}(X)$ of the simplified model was obtained using Eq. (1) changing λ from 0 to 1 in 11 steps and is shown in Fig. 3. As seen from the figure a reasonable behavior is obtained when the folded form is at a lower free energy than the unfolded form. It should be noted that, except for the use of the size-constraint potential given by Eq. (3), no other constraint force was used to drive the conformational change of the 12-mer helix. The main challenge of the present work is not the evaluation of $\Delta g_{\mathrm{sp}}(X)$ but the evaluation of the corresponding free energy of the explicit model. The simulations were averaged over four runs and the corresponding results are given in Fig. 3. Our ability to evaluate the energy of the explicit model was examined using Eq. (11) with U_{sp} as a reference potential (the resulting free energy is $\Delta g_{\mathrm{ep}}^{(11)}$) and by a direct mapping of the explicit all-atom potential (the resulting free energy is $\Delta g_{\mathrm{ep}}^{\mathrm{dir}}$). Both simulations were done with the same conditions as those used for the simplified model. The result of the simulations is shown in Fig. 4 . Although the simulations of $\Delta g_{\mathrm{ep}}^{(11)}$ and $\Delta g_{\mathrm{ep}}^{\mathrm{dir}}$ are not identical they do show similarity that should be improved with better convergence. These results are obviously quite preliminary and are meant mainly to introduce our mapping idea. Nevertheless, these preliminary results illustrate the potential of the present approach in evaluating the free-energy surface of folding processes.

Table 1. Optimized nonbonded parameters[a]

Amino acids	r^0	ε^0
A	2.8	0.05
L	3.5	0.21
I	3.8	0.21
C	3.1	0.10
M	3.8	0.21
P	3.4	0.39
F	4.1	0.16
Y	4.2	0.45
D	3.4	0.21
N	3.3	0.21
T	3.4	0.16
R	4.1	0.39
K	3.8	0.27
G	2.3	0.03
V	3.5	0.16
W	4.4	0.45
E	4.4	0.27
Q	3.7	0.27
H	3.8	0.33
S	2.9	0.10

[a] The simplified nonbonded potential for interaction between protein residues is given by the 8-6 potential, i.e., $\varepsilon_{ij}\left[3\left(\frac{r_{ij}^0}{r_{ij}}\right)^8 - 4\left(\frac{r_{ij}^0}{r_{ij}}\right)^6\right]$, where $\varepsilon_{ij}^0 = \sqrt{\varepsilon_i^0 \varepsilon_j^0}$ and $r_{ij}^0 = \sqrt{r_i^0 r_j^0}$

4 Concluding remarks

The present work develops a new approach for calculations of the folding free energies of all-atom protein models. This new approach is based on using the potential of a simplified protein representation as a reference potential for calculating all-atom free energies. The use of the simplified reference potential allows one to speed up the calculations in a substantial way. Here we take advantage of the fact that the free energy associated with large conformational change can be evaluated in two steps. First we estimate the energetics of moving from the initial to the final regions of the

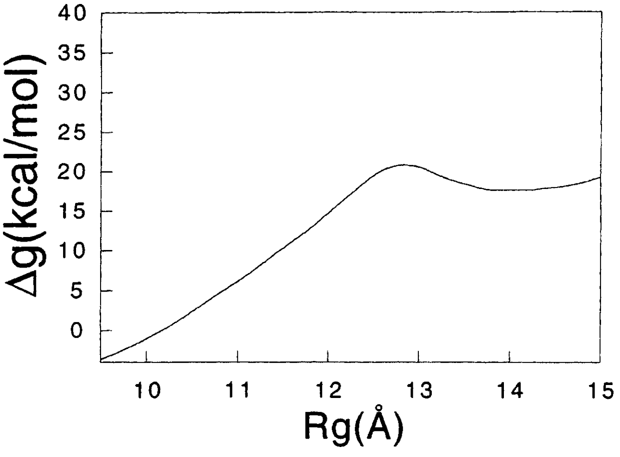

Fig. 3. The free-energy curve $\Delta g_{\mathrm{sp}}(X)$ as a function of R_{g}

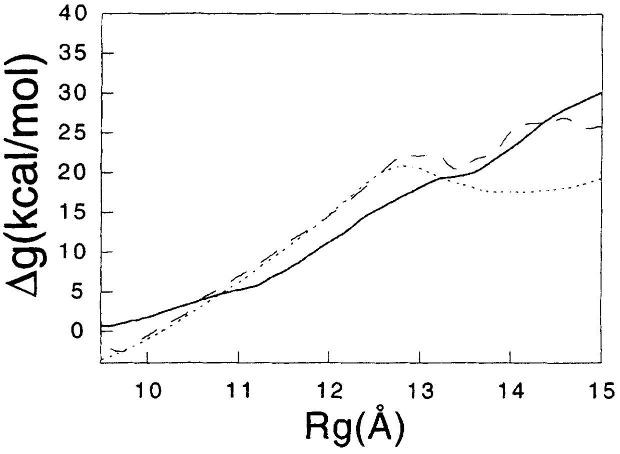

Fig. 4. The free energy curve $\Delta g_{\mathrm{ep}}(X)$ (*solid line*) obtained by Eq. (11) ($\Delta g_{\mathrm{ep}}^{(11)}$) and the corresponding curve (*dashed line*) obtained by direct mapping of the explicit model ($\Delta g_{\mathrm{ep}}^{\mathrm{dir}}$) as a function of R_{g} of the simplified model. The figure also includes the curve $\Delta g_{\mathrm{sp}}(X)$ (*dotted line*). The error range of the calculations is approximately 3 kcal/mol

landscape using the simplified potential. Next we evaluate the free energy associated with moving from the simplified to the explicit potential in the initial and final regions. The ability to map from the simplified to the explicit potential should offer the option to examine the relationship between both approaches. In particular it would be exciting if we find that the change from the simplified to the explicit representation does not lead to a major change in $\Delta g(X)$. This would help examine the range of validity of the simplified representation and to establish to effect of the detailed structure of the sidechains. Our method has been examined in a preliminary way by considering the free-energy curve for the folding of a 12-residue helix. This study demonstrates the main feature of the method and its potential for folding studies. Obviously more extensive studies of the folding of proteins rather than a single helix are essential in order to determine the range of applicability of the method and such studies will be reported in the future.

The present approach can help in providing better parameters for the simplified representation. This can be done by finding parameters that minimize the difference between Δg_{sp} and Δg_{ep}. Furthermore, our model can provide an effective tool in rational drug design; that is, calculations of protein–ligand interactions involve major sampling problems, and require very long simulation times in order to obtain converging results by all-atom models (see discussion in Ref. [24]). This problem can be reduced by evaluating first the binding free energy using a simplified model for both the protein and the ligand, and then calculating the binding free energy of the explicit model using the present approach.

Acknowledgements. This work was supported by NIH grant GM-24492 (A.W.) and NSC grant (J.-K.H).

References

1. Levitt M, Warshel A (1975) Nature 253:694
2. Bryngelson JD, Wolynes PG (1987) Proc Natl Acad Sci USA 84:7524
3. Hinds DA, Levitt M (1992) Proc Natl Acad Sci USA 89:2536
4. Shakhnovich E, Abkevich V, Ptitsyn O (1996) Nature 379:96
5. Dill DA (1990) Biochemistry 29:7133
6. Olszewski KA, Kolinsky A, Skolnick J (1996) Proteins 25:286
7. Daggett V, Levitt M (1993) J Mol Biol 232:600
8. Boczko EM, Brooks CL III (1995) Science 269:393
9. Lazaridis T, Karplus M (1997) Science 278:1928
10. Hwang J-K, Chu ZT, Yadav A, Warshel A (1991) J Phys Chem 95:8445
11. Hwang J-K, Warshel A (1993) J Phys Chem 97:10053
12. Hwang J-K, King G, Creighton S, Warshel A (1988) J Am Chem Soc 110:5297
13. Luzhkov V, Warshel A (1991) J Am Chem Soc 113:4491
14. Muller RP, Warshel A (1995) J Phys Chem 99:17516
15. Hwang J-K, Liao W-F (1995) Protein Eng 8:363
16. Hwang J-K, Warshel A (1996) J Am Chem Soc 118:11745
17. Levitt M (1976) J Mol Biol 104:59
18. Broyden CG (1970) J Inst Math Its App 6:222
19. Fletcher R (1970) Comput J 13:317
20. Goldfarb D (1970) Math Comput 24:23
21. Shanno DF (1970) Math Comput 24:647
22. Lee FS, Chu ZT, Warshel A (1993) J Comput Chem 14:161
23. Warshel A, Russell ST (1984) Q Rev Biophys 17:283
24. Muegge I, Qi PX, Wand AJ, Chu ZT, Warshel A (1997) J Phys Chem B 101:825