Molecular Dynamics Applied to X-ray Structure Refinement

AXEL T. BRUNGER*,† AND PAUL D. ADAMS‡

The Howard Hughes Medical Institute, Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences, and Stanford Synchrotron Radiation Laboratory, Stanford University, 1201 Welch Road, Stanford, California 94305, and Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720

Received July 25, 2001

ABSTRACT

Simulated annealing, in the form of temperature-controlled molecular dynamics, has been successfully applied to macromolecular X-ray structure optimization. The theory and practice of the method are reviewed, and some recent improvements are described.

Introduction

Over the past decade, developments in molecular biology, X-ray diffraction instrumentation, and computational methods have allowed a nearly exponential growth of macromolecular structural studies. In particular, cryoprotection to extend crystal life,[1] the availability of tunable synchrotron sources,[2] high-speed CCD data collection devices,[3] and the ability to incorporate anomalously scattering selenium atoms into proteins have all made structure solution much more efficient.[3] The multiple anomalous diffraction (MAD) method[4] often allows high-quality experimental electron density maps to be obtained. The analysis of the experimental data generally requires sophisticated computational procedures that culminate in refinement and structure validation. This refinement procedure can be formulated as the chemically constrained or restrained nonlinear optimization of a target function, which usually measures the agreement between observed data and data computed from an atomic model. The ultimate goal is to optimize the simultaneous agreement of an atomic model with observed data and with a priori chemical information.

The target function used for this optimization normally depends on several atomic parameters but most importantly on atomic coordinates. The large number of adjustable parameters (typically at least three times the number of atoms in the model) gives rise to a very complicated target function. This in turn produces what is known as the multiple minima problem: the target function contains many local minima in addition to the global minimum. These local minima tend to defeat gradient-descent optimization techniques such as conjugate gradient or least-squares methods.[5] These methods are simply not capable of sampling molecular conformations thoroughly enough to find the most optimal model if the starting one is far from the correct structure.

Simulated annealing is an optimization technique particularly well suited to overcoming the multiple minima problem.[6] Unlike gradient-descent methods, simulated annealing can cross barriers between minima and thus can explore a greater volume of the parameter space to find better models in deeper minima. Following its introduction to crystallographic refinement,[7] there have been major improvements of the original method in four principal areas: the measure of model quality; the search of the parameter space; the target function; the modeling of conformational variability. The combination of improved experimental methods and powerful simulated annealing algorithms has allowed more and more challenging systems to be analyzed, recently culminating in several structures for the ribosome at atomic resolution.[8−10]

For crystallographic refinement, the introduction of cross-validation (the "free" R value) has significantly reduced the danger of overfitting the diffraction data.[11] The complexity of the conformational space has been reduced by the introduction of torsion-angle molecular dynamics,[12] which decreases the number of adjustable parameters that describe a model approximately 10-fold. The target function has been improved by incorporating the concept of maximum likelihood, which takes into account model error, model incompleteness, and errors in the experimental data.[13−15] Finally, the sampling power of simulated annealing can be used for exploring the molecule's conformational space in cases where the molecule undergoes dynamic motion or static disorder through multiconformer models.[16−18]

The Target Function

In essence, macromolecular structure calculation and refinement is a search for the global minimum of a target function

$$E = E_{\text{chem}} + w_{\text{data}} E_{\text{data}} \qquad (1)$$

as a function of the parameters of an atomic model, in particular atomic coordinates. E_{chem} comprises empirical information about chemical interactions; it is a function of all atomic positions, describing covalent (bond lengths, bond angles, torsion angles, chiral centers, and planarity

Dr. Brunger is Investigator in the Howard Hughes Medical Institute and Professor of Molecular and Cellular Physiology, Neurology and Neurological Sciences, and the Stanford Synchrotron Radiation Laboratory at Stanford University. He received his Ph.D. degree from the Technical University of Munich. He held a NATO postdoctoral fellowship and subsequently became a research associate in the Harvard University Department of Chemistry before joining the faculty at Yale University. Dr. Brunger received the Röntgen Prize in biosciences from the University of Würzburg, Würzburg, Germany.

Paul Adams was born in Aldridge, England, in 1966. He received his undergraduate degree in biological sciences and a Ph.D. in biochemistry from Edinburgh University. From there he went to Yale University to work as a Postdoctoral Associate with Axel Brunger. In 1999 he became a Principal Investigator in the Physical Biosciences Division of Lawrence Berkeley Laboratory in California. His research interests focus on the development of new software for structural biology, in particular high-throughput X-ray crystallography.

* To whom correspondence should be addressed. Phone: (650) 736-1031. Fax: (650) 745-1463. E-mail: axel.brunger@stanford.edu.
† Stanford University.
‡ Lawrence Berkeley National Laboratory.

10.1021/ar010034r CCC: $22.00
Published on Web 03/29/2002

of aromatic rings) and nonbonded (intramolecular as well as intermolecular and symmetry-related) interactions. E_{data} describes the difference between observed and calculated data, and w_{data} is a weight appropriately chosen to balance the gradients (with respect to atomic parameters) arising from the two terms.

A Priori Chemical Information

The geometric energy function E_{chem} consists of terms for covalent bonds, bond angles, chirality, planarity, and nonbonded repulsion.[19] The parameters for the covalent terms can be derived from average geometry and root-mean-square (rms) deviations observed in a small-molecule database. Extensive statistical analyses were undertaken for the chemical moieties of proteins[20] and of polynucleotides[21] using the Cambridge Crystallographic Database.[22] Analysis of the ever increasing number of atomic resolution macromolecular crystal structures will no doubt cause some modifications of these parameters in the future.[23−26] It is common to use a purely repulsive quartic function ($E_{repulsive}$) for the nonbonded interactions[19] that are included in E_{chem},

$$E = \sum_{ij}((cR_{ij}^{\min})^n - R_{ij}^n)^m \qquad (2)$$

where R_{ij} is the distance between two atoms i and j, R_{ij}^n is the van der Waals radius for a particular atom pair ij, $c \leq 1$ is a constant that is sometimes used to reduce the radii, and $n = 2$, $m = 2$ or $n = 1$, $m = 4$. In contrast to molecular mechanics force fields, van der Waals attraction and electrostatic interactions are usually not included in structure calculation and refinement. These simplifications are valid since the experimental data contains information that is able to produce atomic conformations consistent with actual nonbonded interactions. In fact, atomic resolution crystal structures can be used to derive parameters for electrostatic energies.[27] Purely repulsive nonbonded interactions are used partly because the calculation is simplified and, therefore, computationally faster. However, the main motivation is to avoid biasing the structure calculation to artifacts that may be present in the force field. In particular, the electrostatic terms are difficult to parametrize. If the experimental diffraction information is insufficient to fully determine the macromolecular structure, use of electrostatic, attractive van der Waals, and simulated solvent interactions can bias the structure toward the theoretical nonbonded model. In this instance it is preferable that the atoms do not attract one another but rather are moved to points of minimal interaction as a result of repulsion.

Geometric energy functions are related to empirical energy functions that were developed for energy-minimization and molecular-dynamics studies of macromolecules.[28] These empirical energy functions were not designed for structure determination and therefore required some modification for use in macromolecular structure refinement.[29−32] Recently, crystallographic simulated annealing refinement was implemented with a purely geometric energy function that uses only covalent energy terms in combination with the repulsive potential described above.[14] This helps to provide uniformity among different crystallographic refinement programs and simplifies the generation of parameters for new chemical compounds.

X-ray Diffraction Data

The conventional form of $E_{X\text{-ray}}$ consists of the crystallographic residual E^{LSQ}, defined as the sum over the squared differences between the observed F_o and calculated F_c structure factor amplitudes for a particular atomic model:

$$E_{X\text{-ray}} = E^{LSQ} = \sum_{hkl}(|F_o| - k|F_c|)^2 \qquad (3)$$

Here hkl are the indices of the reciprocal lattice points of the crystal, F_o and F_c are the observed and calculated structure-factor amplitudes, and k is a relative scale factor.

Reduction of E^{LSQ} can result from improvement in the atomic model but also from an accumulation of systematic errors in the model or fitting noise in the data.[33] The least-squares residual is therefore poorly justified when the model is far away from the correct one or incomplete.[34] An improved target for macromolecular refinement can be obtained using a maximum-likelihood formulation.[13,35−38] The goal of the maximum-likelihood method is to determine the probability of making a measurement, given the model and estimates of the model's errors and those of the measured intensities. The effects of model errors (incorrectly placed and missing atoms) on the calculated structure factors are first quantified with σ_A values, which correspond roughly to the fraction of each structure factor that is expected to be correct. However, overfitting of the diffraction data causes the model bias to be underestimated and undercorrected in the σ_A values. The effect of this overfitting can be reduced by cross-validating σ_A values, i.e., by computing them from a randomly selected test set that is excluded from the summation[39,40] on the right-hand side of eq 3. The expected values of $\langle F_o \rangle$ and the corresponding variance (σ_{ML}^2) are derived from σ_A, the observed F_o, and calculated F_c. These quantities can be readily incorporated into a maximum-likelihood target function:[13]

$$E_{X\text{-ray}} = E_{ML} = \sum_{hkl \in \text{workingset}} \left(\frac{1}{\sigma_{ML}^2}\right)(|F_o| - \langle |F_o| \rangle)^2 \qquad (4)$$

To achieve an improvement over the least-squares residual (eq 3), cross-validation was found to be essential[14] for the computation of σ_A and its derived quantities in eq 4.

For many crystal structures, some initial experimental phase information is available from either isomorphous heavy-atom replacement, single- or multiwavelength anomalous diffraction methods. These phases represent additional observations that can be incorporated in the refinement target. The maximum likelihood formulation

naturally extends itself to incorporation of this information.[15,41] Tests have shown that the addition of experimental phase information, including single-isomorphous replacement (SIR) or single-wavelength anomalous dispersion (SAD), greatly improves the results of refinement.[15,40] It should be noted that with advances in synchrotron X-ray radiation instrumentation and density modification it is now possible to use SAD phasing to solve a crystal structure in favorable cases.[42]

Pannu and Read[13] have developed an efficient Gaussian approximation for the case of structure factor amplitudes with no prior phase information, termed the MLF target function. In the limit of a perfect model MLF reduces to the traditional least-squares residual (eq 3) with $1/\sigma^2$ weighting. In the case where prior phase information is included, the integration over the phase angles is carried out numerically and is termed the MLHL target.[15] A maximum likelihood function that expresses the probability distributions in terms of observed intensities has also been developed and is termed MLI.[13]

Additional Information

Additional constraints or restraints may be used to effectively improve the ratio of observations to parameters. For example, atoms can be grouped so that they move as rigid bodies during refinement or bond lengths and bond angles can be kept fixed.[12,43,44] The existence of noncrystallographic symmetry can be used to average over equivalent molecules and thereby to reduce noise in the diffraction data.[30]

Weighting

The weight w_{data} (eq 1) balances the forces arising from E_{data} and E_{chem}. The choice of w_{data} can be critical: if w_{data} is too large, the refined structure will show unphysical deviations from ideal geometry; if w_{data} is too small, the refined structure will not satisfy the observed data. Automated protocols to provide initial estimates for optimal weighting have been developed.[14,29] However, independent information must be used (e.g. cross-validation) to objectively obtain the best possible weight for the X-ray diffraction data.[11]

Searching Conformational Space

Annealing denotes a physical process wherein a solid is heated until all particles randomly arrange themselves in a liquid phase and then is cooled slowly so that all particles arrange themselves in the lowest energy state. By formally defining the target E (eq 1) to be the equivalent of the potential energy of the system, one can simulate the annealing process.[6] There is no guarantee that simulated annealing will find the global minimum (except in the case of an infinitely long search).[45] Compared to conjugate-gradient minimization where search directions must follow the gradient, simulated annealing achieves more optimal solutions by allowing motion against the gradient.[6] The likelihood of uphill motion is determined by a control parameter referred to as temperature. The higher the temperature, the more likely it is that simulated annealing will overcome barriers. It should be noted that the simulated annealing temperature normally has no physical meaning and merely determines the likelihood of overcoming barriers of the target function.

The simulated annealing algorithm requires a generation mechanism to create a Boltzmann distribution at a given temperature T. Simulated annealing also requires an annealing schedule; that is, a sequence of temperatures $T_1 > T_2 > ... > T_n$ at which the Boltzmann distribution is computed. Implementations of the generation mechanism differ in the way they transition from one set of parameters to another that is consistent with the Boltzmann distribution at a given temperature. The two most widely used generation mechanisms are Metropolis Monte Carlo[46] and molecular dynamics[47] simulations. For X-ray crystallographic refinement, molecular dynamics has proved extremely successful[3] whereas Monte Carlo methods have yet to be shown to be effective.

Molecular Dynamics

A suitably chosen set of atomic parameters can be viewed as generalized coordinates that are propagated in time by the classical (Hamilton) equations of motion.[48] If the generalized coordinates represent the x, y, z positions of the atoms of a molecule, the Hamilton equations of motion reduce to the more familiar Newton's second law:

$$m_i \frac{\partial^2 \vec{r}_i}{\partial t^2} = -\nabla_i E \qquad (5)$$

The quantities m_i and r_i are respectively the mass and coordinates of atom i, and E is given by eq 1. The solution of the partial differential equations (eq 5) is achieved numerically using finite-difference methods.[47] This approach is referred to as molecular dynamics.

Initial velocities for the integration of eq 5 are usually assigned randomly from a Maxwell distribution at the appropriate temperature. Assignment of different initial velocities will produce a somewhat different structure after simulated annealing. By performing several refinements with different initial velocities, one can therefore improve the chances of success of simulated annealing refinement. Furthermore, this improved sampling can be used to study discrete disorder and conformational variability (see below).

Although Cartesian (i.e., flexible bond lengths and bond angles) molecular dynamics places restraints on bond lengths and bond angles (through E_{chem}, eq 1), one might want to implement these restrictions as constraints, i.e., fixed bond lengths and bond angles.[43] This is supported by the observation that the deviations from ideal bond lengths and bond angles are usually small in X-ray crystal structures. Indeed, fixed-length constraints have been applied to structure calculation by least-squares or conjugate-gradient minimization.[43] It is only recently, however, that efficient and robust algorithms have become available for molecular dynamics in torsion-angle space.[49–52]

Using an approach that retains the Cartesian-coordinate formulation of the target function and its derivatives with respect to atomic coordinates makes calculations remains relatively straightforward and topology independent.[12] In this formulation, however, the expression for the acceleration becomes a function of positions and velocities. Iterative equations of motion for constrained dynamics in this formulation can be derived and solved by finite difference methods.[53] This method is numerically very robust and has a significantly increased radius of convergence in crystallographic refinement compared to Cartesian molecular dynamics.[12]

Temperature Control

Simulated annealing requires the control of the temperature during molecular dynamics. The current temperature of the simulation (T_{curr}) is computed from the kinetic energy

$$E_{\mathrm{kin}} = \sum_{i}^{n\ \mathrm{atoms}} \frac{1}{2} m_i \left(\frac{\partial r_i}{\partial t}\right)^2 \qquad (6)$$

of the molecular dynamics simulation,

$$T_{\mathrm{curr}} = \frac{2E_{\mathrm{kin}}}{3nk_{\mathrm{b}}} \qquad (7)$$

Here n is the number of degrees of freedom and k_{b} is Boltzmann's constant. One commonly used approach to control the temperature of the simulation consists of coupling the equations of motion to a heat bath. A friction term (γ_i) to control the temperature

$$-m_i\gamma_i v_i\left(1 - \left(\frac{T}{T_{\mathrm{curr}}}\right)\right) \qquad (8)$$

can be added to the right-hand side of eq 5, where v_i are the velocities of the atoms.[54] This method generalizes the concept of friction by allowing a negative friction coefficient and by determining the friction coefficient and its sign by the ratio of the current simulation temperature to the target temperature T_{curr}.

Why Does Simulated Annealing Work?

The goal of any optimization problem is to find the global minimum of a target function. In the case of macromolecular structure calculation and refinement, one searches for the conformation or conformations of the molecule that best fit the experimental data and that simultaneously maintain reasonable covalent and noncovalent interactions. Simulated annealing refinement has a much larger radius of convergence than conjugate-gradient minimization (see below). It must therefore be able to find a lower minimum of the target E (eq 1) than the local minimum found by simply moving along the negative gradient of E. Paradoxically, the very reasons that make simulated annealing such a powerful refinement technique (the ability to overcome barriers in the target energy function) would seem to prevent it from working at all. If it crosses barriers so easily, what allows it to stay in the vicinity of the global minimum?

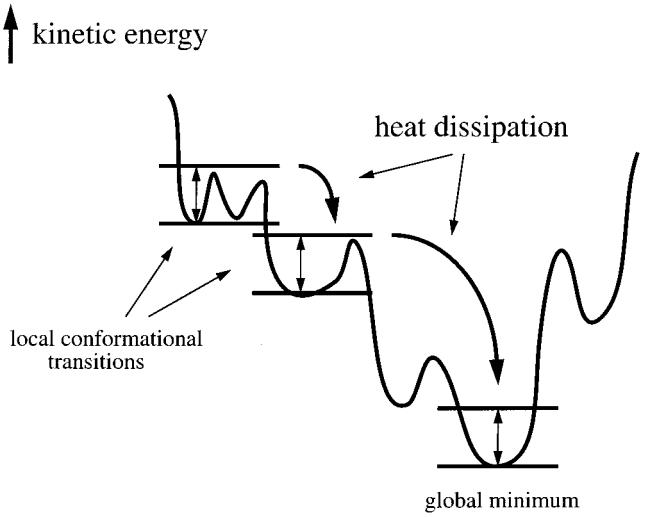

FIGURE 1. Schematic explanation of molecular dynamics based simulated annealing. The kinetic energy of the system allows local conformational transitions with barriers smaller than the kinetic energy. If a larger drop in energy is encountered, the excess kinetic energy is dissipated through the friction term (eq 8). It is thus unlikely that the system can climb out of the global minimum once it has reached it.

It is most easy to visualize this property of simulated annealing in the case of molecular dynamics. When a fixed temperature is specified, the system essentially gains a certain inertia that allows it to cross energy barriers of the corresponding target function (eq 7). The target temperature must be large enough to overcome smaller barriers (e.g., Figure 1) but low enough to ensure that the system will not "climb out" out of the global minimum if it manages to arrive there. While temperature itself is a global parameter of the system, temperature fluctuations arise principally from local conformational transitions—for example from an amino acid side chain falling into the correct orientation. These local changes tend to lower the value of the target E, thus increasing the kinetic energy and, hence, the temperature of the system. Once the temperature coupling (eq 8) has removed this excess kinetic energy through heat dissipation, the reverse transition is very unlikely, since it would require a localized increase in kinetic energy where the conformational change occurred in the first place. Temperature coupling maintains a sufficient amount of kinetic energy to allow local conformational corrections but does not supply enough to allow escape from the global minimum. This explains the observation that on average the agreement with the experimental data will improve rather than worsen with simulated annealing.

Practical Considerations

As Figure 1 illustrates, the simulation temperature needs to be high enough to allow conformational transitions but not too high to avoid moving too far away from the initial structure. The optimum temperature for a given starting structure is a matter of trial and error. We empirically

determined starting temperatures for a variety of simulated annealing protocols[14,55] that should work for the average case. However, it might be worth trying a different temperature if a particularly difficult refinement problem is encountered. In particular, significantly higher temperatures are attainable using torsion-angle molecular dynamics. Note that each simulated annealing refinement run is subject to chance by using a random number generator to generate the initial velocities. Thus, multiple refinements must be run if systematic trends resulting from changes of certain parameters of the annealing schedule are to be studied. The best structure(s) among a set of refinements using different initial velocities and/or temperatures should be taken for further refinement or averaging (see below).

The annealing schedule employed can in principle be any function of the simulation step (or time domain). The two most commonly used protocols are linear slow cooling or constant temperature followed by quenching. A slight advantage is obtained with slow cooling.[31] The duration of the annealing schedule is another parameter. Too short a protocol does not allow sufficient sampling of conformational space. Too long a protocol may waste computer time since it is more efficient to run multiple trials as opposed to one long refinement protocol (unpublished results).

Crystallographic Refinement

In the crystallographic case, the limited radius of convergence of refinement arises not only from the high dimensionality of the parameter space but also from the crystallographic phase problem. For new crystal structures, initial electron density maps must be computed from a combination of observed diffraction amplitudes and experimental phases where the latter are typically of poorer quality and lower resolution than the former. A different problem arises when structures are solved by molecular replacement, which uses a homologous structure as a search model.[56,57] In this case the resulting electron density maps can be severely "model-biased"; that is, they seem to confirm the existence of the search model without providing clear evidence of actual differences between it and the true crystal structure. In either case, initial atomic models usually require extensive refinement.

Many examples have shown that simulated annealing refinement starting from initial models (obtained by standard crystallographic techniques) produces significantly better final models compared to those produced by least-squares or conjugate-gradient minimization. In a realistic test case,[40] a series of models for the aspartic proteinase penicillopepsin was generated from homologous structures present in the Protein Data Bank. The sequence identity among these structures ranged from 100% to 25%, thus providing a set of models with increasing coordinate error compared to the refined structure of penicillopepsin. These models, after truncation of all residues to alanine, were all used as search models in molecular replacement against the native penicillopepsin diffraction data. In all cases the correct placement of the model in the penicillopepsin unit cell was found.

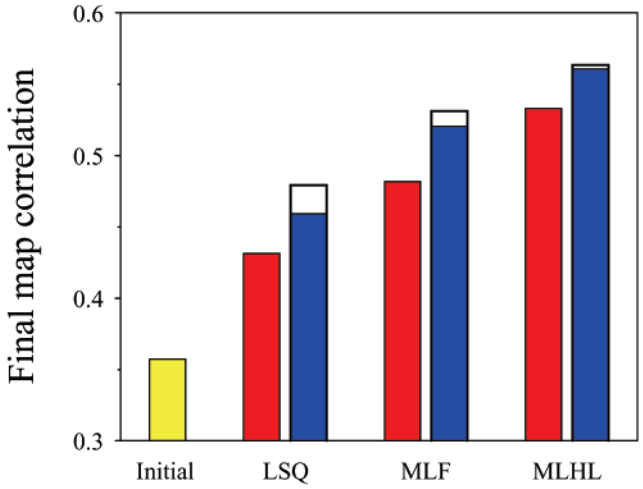

FIGURE 2. Simulated annealing (blue bars) produces better models than extensive conjugate gradient minimization (red bars). Map correlation coefficients were computed before and after refinement against the native penicillopepsin diffraction data[66] for the polyalanine model derived from *Mucor pusillus* pepsin.[67] Correlation coefficients are between σ_A-weighted maps calculated from each model and from the published penicillopepsin structure. The observed penicillopepsin diffraction data was in space group *C*2 with cell dimensions $a = 97.37$ Å, $b = 46.64$ Å, $c = 65.47$ Å, and $\beta = 115.4°$. All refinements were carried out using diffraction data from the lowest resolution limit of 22.0 Å up to 2.0 Å. The MLHL refinements used single-isomorphous phases from a $K_3UO_2F_5$ derivative of the penicillopepsin crystal structure, which covered a resolution range of 22.0 to 2.8 Å. Simulated annealing refinements were repeated 5 times with different initial velocities. The numerical averages of the map correlation coefficients for the 5 refinements are shown as the blue bars. The best map correlation coefficients from simulated annealing are shown as the white bars.

Both conjugate gradient minimization and simulated annealing were carried out to compare the performance of LSQ (the least-squares residual), MLF (the maximum likelihood target using amplitudes), and MLHL (the maximum likelihood target using amplitudes and experimental phase information). In the latter case, phases from single-isomorphous replacement were used. A very large number of conjugate gradient cycles were carried out to make the computational requirements equivalent for both minimization and simulated annealing. The conjugate gradient minimizations were converged; i.e., there was no change when further cycles were carried out.

For a given target function, simulated annealing always outperformed minimization (Figure 2). For a given starting model, the maximum likelihood targets outperformed the least-squares residual target for both minimization and simulated annealing, producing models with lower phase errors and higher map correlation coefficients when compared to the published penicillopepsin crystal structure (Figure 2). This improvement is also illustrated in the σ_A-weighted electron density maps obtained from the refined models (Figure 3). The incorporation of experimental phase information further improved the refine-

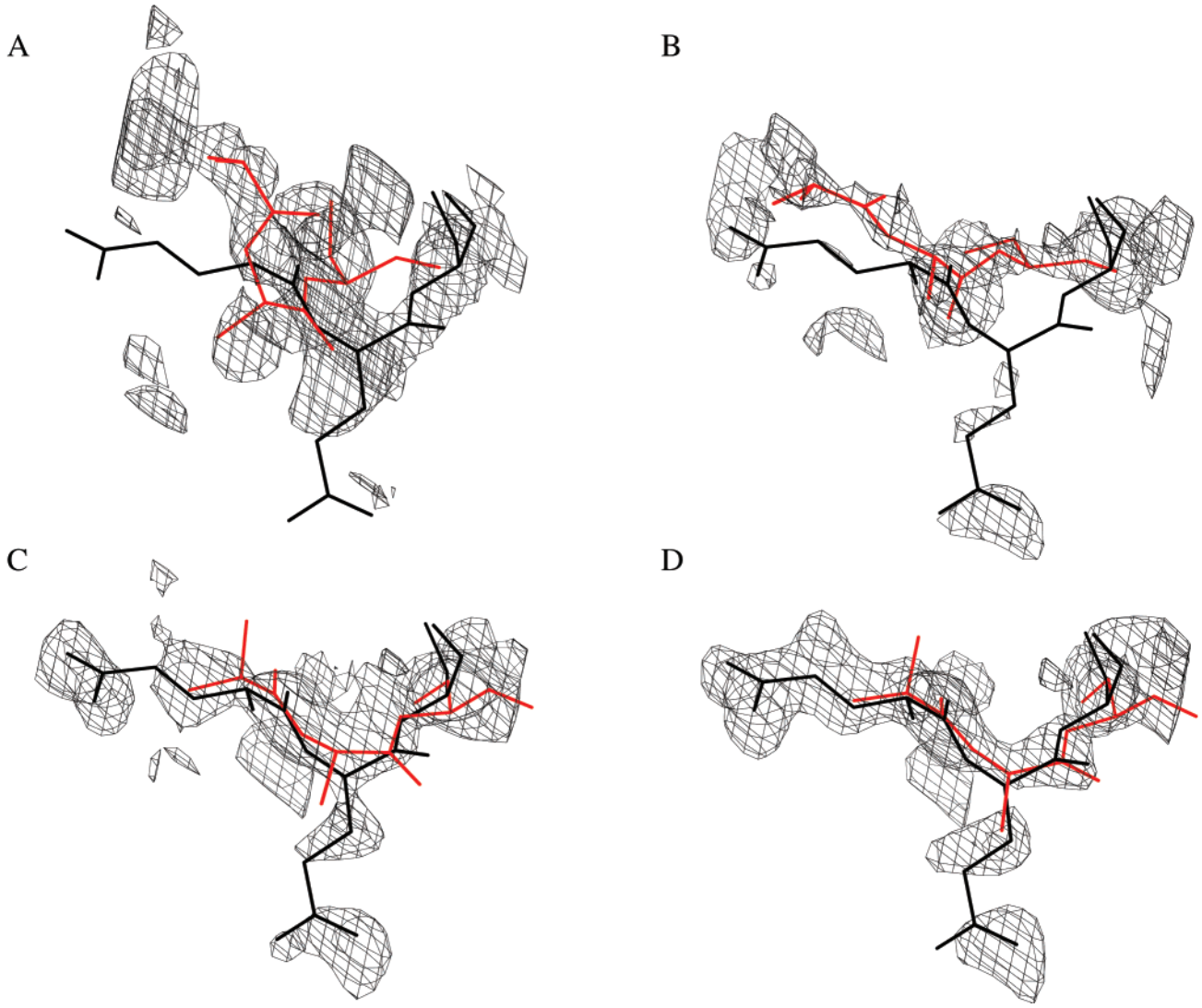

FIGURE 3. Maximum likelihood targets significantly decrease model bias in simulated annealing refinement. σ_A-weighted electron density maps contoured at 1.25 σ for models from simulated annealing refinement with different targets. Residues 50−52 are shown, with the published penicillopepsin crystal structure[66] in black and the model with the lowest free R value from 5 independent refinements in red. Key: (a) initial electron-density map prior to refinement; (b) after refinement with the LSQ target; (c) after refinement with the MLF target; (d) after refinement with the MLHL target.

ment significantly despite the ambiguity in the SIR phase probability distributions. Thus, the most efficient refinement will make use of torsion angle dynamics simulated annealing and prior phase information in the MLHL maximum likelihood target function.

Cross-validation is essential in the calculation of the maximum likelihood target.[13,14,39] Maximum-likelihood refinement without cross-validation gives much poorer results, as indicated by higher free R values, higher $R_{free} - R$ differences, and larger phase errors. It should be noted that the normal R value usually increases upon using the cross-validated maximum likelihood formulation. This is a consequence of the reduction of overfitting by this method.

Simulated annealing refinement is most useful when the initial model is relatively crude. Given a well-refined model, it offers little advantage over conventional methods for automatically improving the model, with the important exception of reducing model bias in annealed omit maps.[58] Furthermore, simulated annealing refinement of a final model can provide information about the accuracy and conformational variability of the refined structure (see below).

Averaging of Independently Refined Structures

As mentioned above, multiple simulated annealing refinements will generally produce somewhat different structures, some of which may be better (as assessed, for example, in terms of the free R value) than others. This approach offers several advantages. First, a more optimum structure can be obtained from multiple trials as opposed to a single simulated annealing calculation. Second, each member of the family of refined structures may be better in different regions of the molecule. Thus, by examining the ensemble during model-building, one may gain insights into possible local conformations of the molecule. Third, the structure factors of all structures of the family

may be averaged. This averaging will reduce the effect of local errors (noise) that are presumably different in each member of the family.

Torsion-angle molecular dynamics simulated annealing with the maximum likelihood target (eq 4) performed on human heterogeneous ribonucleoprotein A1, hnRNP[59] showed that averaging produced the least model-biased map (as indicated by the lowest free R value and the lowest $R_{free} - R$ difference) with the polypeptide backbone being completely connected.[60] This example is another demonstration that cross-validation of the R value is essential for assessing model correctness[11] since the normal R value decreases with increasing model-bias of the electron density maps whereas the free R value shows the correct behavior.

Ensemble Models

In cases of conformational variability or discrete disorder, there is not a single correct solution to the optimization problem eq 1. Rather, the X-ray diffraction data represent a spatial and temporal average over all conformations that are assumed by the molecule. Ensembles of structures, which are simultaneously refined against the observed data, may thus be a more appropriate description of the data. This has been used for some time in X-ray crystallography when alternate conformations are modeled locally. Alternate conformations can be generalized to global conformations;[16,17,61] i.e.; the model is duplicated n-fold, the corresponding calculated structure factors are added and refined simultaneously against the observed X-ray diffraction data, and each member of the family is chemically "invisible" to all other members. The number n can be determined by cross-validation.[17,18]

An advantage of a multiconformer model is that it directly incorporates many possible types of disorder and motion (global disorder, local side chain disorder, local wagging and rocking motions), although it is not generally possible to distinguish between the static disorder and motion with data from a single experiment. Furthermore, a multiconformer model can be used to automatically detect the most variable regions of the molecule by inspecting the atomic root-mean-square difference around the mean as a function of residue number. Thermal factors of single conformer models may sometimes be misleading by underestimating the degree of motion or disorder,[62] and thus, the multiple-conformer model can be a more faithful representation of the diffraction data. However, it should be noted that when very high-resolution experimental data (d_{min} of approximately 1.0 Å) are available, the use of anisotropic thermal factors is seen to give a better fit to the data.[63] A disadvantage of the multiconformer model is that it introduces many more degrees of freedom. However, cross-validated maximum-likelihood refinement can address this problem. For example, the R_{free} and R values were 0.239 and 0.237 for a single conformer refinement and 0.231 and 0.230, respectively, for a four-conformer refinement at 50−1.7 Å resolution data of a fragment of mannose-binding protein A[18] illustrating that introduction of multiple conformers did not increase the amount of overfitting compared to the single-conformer case (unpublished results).

Although there are some similarities between averaging individually refined structures and multiconformer models, there are also fundamental differences. For example, in the case of X-ray crystallography, averaging seeks to improve the calculated electron density map by averaging out the noise present in the individual models, each of which is still a good representation of the diffraction data. This method is most useful at the early stages of refinement when the model still contains errors. In contrast, multiconformer refinement seeks to create an ensemble of structures at the final stages of refinement which, taken together, best represent the data. It should be noted that each individual conformer of the ensemble does not necessarily remain a good description of the data since the whole ensemble is refined against the data. Clearly, this method requires high-quality data and a high observation-to-parameter ratio.

Conclusions

Simulated annealing has improved the efficiency of macromolecular structure calculation and refinement significantly in X-ray crystallography. A case in point is the combination of torsion angle molecular dynamics with a cross-validated maximum likelihood target, which interact synergistically to produce less model bias than any other method to date. The combined method also increases the radius of convergence allowing the refinement of poor initial models, e.g., those obtained by weak molecular replacement solutions.[12,41] However, simulated annealing refinement alone is still insufficient to refine a structure automatically without human intervention. For example, crystallographic refinement using simulated annealing typically cannot correct mainchain tracing errors such as register shifts. Fully automatic structure determination probably requires significant new algorithmic developments.[64]

Molecular dynamics can also be used to provide new physical insights into molecular function, which may depend on conformational variability. The sampling characteristics of simulated annealing allow the generation of multiconformer models that can represent molecular motion and discrete disorder, especially when combined with the acquisition of high-quality data.[18] Simulated annealing is thus also a stepping stone toward development of improved models of macromolecules.

Many of the recent computational developments discussed in this review are available in the program Crystallography & NMR System (Brunger, Adams, Clore, DeLano, Gros, Grosse-Kunstleve, Jiang, Kuszewski, Nilges, Pannu, Read, Rice, Simonson, and Warren; URL: http://cns.csb.yale.edu).[65] *This work was funded in part by the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.*

References

(1) Garman, E. Cool data: quantity AND quality. *Acta Crystallogr.* **1999**, *D55*, 1641−1653.

(2) Walsh, M. A.; Evans, G.; Sanishvili, R.; Dementieva, I.; Joachimiak, A. MAD data collection-current trends. *Acta Crystallogr.* **1999**, *D55*, 1726−1732.
(3) Walsh, M. A.; Dementieva, I.; Evans, G.; Sanishvili, R.; Joachimiak, A. Taking MAD to the extreme: Ultrafast protein structure determination. *Acta Crystallogr.* **1999**, *D55*, 1168−1173.
(4) Hendrickson, W. A. Determination of Macromolecular Structures from Anomalous Diffraction of Synchrotron Radiation. *Science* **1991**, *254*, 51−58.
(5) Press, W. H.; Flannery, B. P.; Teukolosky, S. A.; Vetterling, W. T. *Numerical Recipes*; Cambridge University Press: Cambridge, U.K., 1986; pp 498−546.
(6) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P., Jr. Optimization by simulated annealing. *Science* **1983**, *220*, 671−680.
(7) Brunger, A. T.; Kuriyan, J.; Karplus, M. Crystallographic R factor refinement by molecular dynamics. *Science* **1987**, *235*, 458−460.
(8) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **2000**, *289*, 905−920.
(9) Wimberly, B. T.; Brodersen, D. E.; Clemons, W. M., Jr.; Morgan-Warren, R. J.; Carter, A. P.; Vonrhein, C.; Hartsch, T.; Ramakrishnan, V. Structure of the 30S ribosomal subunit. *Nature* **2000**, *407*, 327−339.
(10) Schluenzen, F.; Tocilj, A.; Zarivach, R.; Harms, J.; Gluehmann, M.; Janell, D.; Bashan, A.; Bartels, H.; Agmon, I.; Franceschi, F.; Yonath, A. Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* **2000**, *102*, 615−623.
(11) Brunger, A. T. The Free *R* value: a Novel Statistical Quantity for Assessing the Accuracy of Crystal Structures. *Nature* **1992**, *355*, 472−474.
(12) Rice, L. M.; Brunger, A. T. Torsion Angle Dynamics: Reduced Variable Conformational Sampling Enhances Crystallographic Structure Refinement. *Proteins: Struct., Funct., Genet.* **1994**, *19*, 277−290.
(13) Pannu, N. S.; Read, R. J. Improved Structure Refinement Through Maximum Likelihood. *Acta Crystallogr.* **1996**, *A52*, 659−668.
(14) Adams, P. D.; Pannu, N. S.; Read, R. J.; Brunger, A. T. Cross-validated Maximum Likelihood Enhances Crystallographic Simulated Annealing Refinement. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 5018−5023.
(15) Pannu, N. S.; Murshudov, G. N.; Dodson, E. J.; Read, R. J. Incorporation of prior phase information strengthens maximum likelihood structural refinement. *Acta Crystallogr.* **1998**, *D54*, 1285−1294.
(16) Kuriyan, J.; Ösapay, K.; Burley, S. K.; Brunger, A. T.; Hendrickson, W. A.; Karplus, M. Exploration of Disorder in Protein Structures by X-ray Restrained Molecular Dynamics. *Proteins* **1991**, *10*, 340−358.
(17) Burling, F. T.; Brunger, A. T. Thermal Motion and Conformational Disorder in Protein Crystal Structures: Comparison of Multi-Conformer and Time-Averaging Models. *Isr. J. Chem.* **1994**, *34*, 165−175.
(18) Burling, F. T.; Weis, W. I.; Flaherty, K. M.; Brunger, A. T. Direct Observation of Protein Solvation and Discrete Disorder With Experimental Crystallographic Phases. *Science* **1996**, *271*, 72−77.
(19) Hendrickson, W. A. Stereochemically restrained refinement of macromolecular structures. *Methods Enzymol.* **1985**, *115*, 252−270.
(20) Engh, R. A.; Huber, R. Accurate bond and angle parameters for X-ray structure refinement. *Acta Crystallogr.* **1991**, *A47*, 392−400.
(21) Parkinson, G.; Vojtechovsky, J.; Clowney, L.; Brunger, A. T.; Berman, H. M. New Parameters for the Refinement of Nucleic Acid Containing Structures. *Acta Crystallogr.* **1996**, *D52*, 57−64.
(22) Allen, F. H.; Kennard, O.; Taylor, R. Systematic Analysis of Structural Data as a Research Technique in Organic Chemistry. *Acc. Chem. Res.* **1983**, *16*, 146−153.
(23) Dauter, Z.; Lamzin, V. S.; Wilson, K. S. Proteins at atomic resolution. *Curr. Opin. Struct. Biol.* **1995**, *5*, 784−790.
(24) Stec, B.; Zhou, R.; Teeter, M. M. Full-matrix refinement of the protein crambin at 0.83 Å and 130 K. *Acta Crystallogr.* **1995**, *D51*, 663−681.
(25) Sevcik, J.; Dauter, Z.; Lamzin, V. S.; Wilson, K. S. Ribonuclease from Streptomyces aureofaciens at atomic resolution. *Acta Crystallogr.* **1996**, *D52*, 327−344.
(26) Vlassi, M.; Dauter, Z.; Wilson, K. S.; Kokkinidis, M. Structural parameters for proteins derived from the atomic resolution (1.09 Å) structure of a designed variant of the *colE1* ROP protein. *Acta Crystallogr.* **1998**, *D54*, 1245−1260.
(27) Pearlman, D. A.; Kim, S.-H. Atomic charges for DNA constituents derived from single-crystal X-ray diffraction data. *J. Mol. Biol.* **1990**, *211*, 171−187.
(28) Karplus, M.; Petsko, G. A. Molecular dynamics simulations in biology. *Nature* **1990**, *347*, 631−639.
(29) Brunger, A. T.; Karplus, M.; Petsko, G. A. Crystallographic Refinement by Simulated Annealing: Application to a 1.5 Å Resolution Structure of Crambin. *Acta Crystallogr.* **1989**, *A45*, 50−61.
(30) Weis, W. I.; Brunger, A. T.; Skehel, J. J.; Wiley, D. C. Refinement of the Influenza Virus Haemagglutinin by Simulated Annealing. *J. Mol. Biol.* **1989**, *212*, 737−761.
(31) Brunger, A. T.; Krukowski, A.; Erickson, J. Slow-Cooling Protocols for Crystallographic Refinement by Simulated Annealing. *Acta Crystallogr.* **1990**, *A46*, 585−593.
(32) Fujinaga, M.; Gros, P.; van Gunsteren, W. F. Testing the method of crystallographic refinement using molecular dynamics. *J. Appl. Crystallogr.* **1989**, *22*, 1−8.
(33) Silva, A. M.; Rossmann, M. G. The refinement of southern bean mosaic virus in reciprocal space. *Acta Crystallogr.* **1985**, *B41*, 147−157.
(34) Read, R. J. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr.* **1986**, *A42*, 140−149.
(35) Read, R. J. Structure-factor probabilities for related structures. *Acta Crystallogr.* **1990**, *A46*, 900−912.
(36) Bricogne, G. A multisolution method of phase determination by combined maximization of entropy and likelihood. III. Extension to powder diffraction data. *Acta Crystallogr.* **1991**, *A47*, 803−829.
(37) Bricogne, G. Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. *Acta Crystallogr.* **1993**, *D49*, 37−60.
(38) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr.* **1997**, *D53*, 240−255.
(39) Kleywegt, G. J.; Brunger, A. T. Cross-validation in crystallography: practice and applications. *Structure* **1996**, *4*, 897−904.
(40) Adams, P. D.; Pannu, N. S.; Read, R. J.; Brunger, A. T. Extending the limits of molecular replacement through combined simulated annealing and maximum likelihood refinement. *Acta Crystallogr.* **1999**, *D55*, 181−190.
(41) Bricogne, G. Bayesian statistical viewpoints on structure determination: basic concepts and examples. *Methods Enzymol.* **1997**, *276*, 361−423.
(42) Rice, L. M.; Earnest, T. N.; Brunger, A. T. Single wavelength anomalous diffraction phasing revisited. *Acta Crystallogr.* **2000**, *D56*, 1413−1420.
(43) Diamond, R. A real-space refinement procedure for proteins. *Acta Crystallogr.* **1971**, *A27*, 436−452.
(44) Sussman, J. L.; Holbrook, S. R.; Church, G. M.; Kim, S.-H. Structure-factor least-squares refinement procedure for macromolecular structure using constrained and restrained parameters. *Acta Crystallogr.* **1977**, *A33*, 800−804.
(45) Laarhoven, P. J. M.; Aarts, E. H. L. *Simulated Annealing: Theory and Applications*; D. Reidel Publishing Co.: Dordrecht, The Netherlands, 1987.
(46) Metropolis, N.; Rosenbluth, M.; Rosenbluth, A.; Teller, A.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(47) Verlet, L. Computer Experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **1967**, *159*, 98−105.
(48) Goldstein, H. *Classical Mechanics*, 2nd ed.; Addison-Wesley Pub. Co.: Reading, MA, 1980.
(49) Bae, D.-S.; Haug, E. J. A recursive formulation for constrained mechanical system dynamics: Part I. Open loop systems. *Mech. Struct. Mach.* **1987**, *15*, 359−382.
(50) Bae, D.-S.; Haug, E. J. A recursive formulation for constrained mechanical system dynamics: Part II. Closed loop systems. *Mech. Struct. Mach.* **1988**, *15*, 481−506.
(51) Jain, A.; Vaidehi, N.; Rodriguez, G. A Fast Recursive Algorithm for Molecular Dynamics Simulation. *J. Comput. Phys.* **1983**, *106*, 258−68.
(52) Mathiowetz, A. M.; Jain, A.; Karasawa, N.; Goddard, W. A. Protein Simulations Using Techniques Suitable for Very Large Systems: The Cell Mulipole Method for Nonbond Interactions and the Newton-Euler Inverse Mass Operator Method for Internal Coordinate Dynamics. *Proteins: Struct., Funct., Genet.* **1994**, *20*, 227−247.
(53) Abramowitz, M.; Stegun, I. *Handbook of Mathematical Functions;* Applied Mathematics Series Vol. 55; Dover Publications: New York, 1968; p 896.
(54) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(55) Brunger, A. T. Crystallographic refinement by simulated annealing: application to a 2.8 Å resolution structure of aspartate aminotransferase. *J. Mol. Biol*. **1988**, *203*, 803−816.
(56) Hoppe, W. Die Faltmolekülmethode-eine neue Methode zur Bestimmung der Kristallstruktur bei Ganz oder Teilweise bekannter Molekülstruktur. *Acta Crystallogr*. **1957**, *10*, 750−751.
(57) Rossmann, M. G.; Blow, D. M. The detection of subunits within the crystallographic asymmetric unit. *Acta Crystallogr*. **1962**, *A15*, 24−51.
(58) Hodel, A.; Kim, S.-H.; Brunger, A. T. Model bias in macromolecular crystal structures. *Acta Crystallogr*. **1992**, *A48*, 851−859.
(59) Shamoo, Y.; Krueger, U.; Rice, L. M.; Williams, K. R.; Steitz, T. A. Crystal structure of the Two RNA-Binding Domains of Human hnRNP A1 at 1.75 Å resolution. *Nat. Struct. Biol.* **1997**, *3*, 215−222.
(60) Rice, L. M.; Shamoo, Y.; Brunger, A. T. Phase improvement by multi-start simulated annealing refinement and structure factor averaging. *J. Appl. Crystallogr*. **1998**, *31*, 798−805.
(61) Gros, P.; van Gunsteren, W. F.; Hol, W. G. J. Inclusion of Thermal Motion in Crystallographic Structures by Restrained Molecular Dynamics. *Science* **1990**, *249*, 1149−1152.
(62) Kuriyan, J.; Petsko, G. A.; Levy, R. M.; Karplus, M. Effect of Anisotropy and Anharmonicity on Protein Crystallographic Refinement. *J. Mol. Biol*. **1986**, *190*, 227−254.
(63) Wilson, M. A.; Brunger A. T. The 1.0 Å crystal structure of Ca^{2+} bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity. *J. Mol. Biol.* **2000**, *301*, 1237−1256.
(64) Adams, P. D.; Grosse-Kunstleve, R. W. Recent developments in software for automation of crystallographic macromolecular structure determination. *Curr. Opin. Struct. Biol*. **2000**, *10*, 564−568.
(65) Brunger, A. T.; Adams, P. D.; Clore, G. M.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J.-S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Crystallography & NMR system (CNS): A new software system for macromolecular structure determination. *Acta Crystallogr*. **1998**, *D54*, 905−921.
(66) Hsu, I. N.; Delbaere, L. T. J.; James, M. N. G.; Hoffman, T. Penicillopepsin from *Penicillium janthinellum* crystal structure at 2.8 Å and sequence homology with porcine pepsin. *Nature* **1977**, *266*, 140−145.
(67) Newman, M.; Watson, F.; Roychowdhury, P.; Jones, H.; Badasso, M.; Cleasby, A.; Wood, S. P.; Tickle, I. J.; Blundell, T. L. X-ray analyses of aspartic proteinases. V. Structure and refinement at 2.0 Å resolution of the aspartic proteinase from *Mucor pusillus*. *J. Mol. Biol*. **1993**, *230*, 260−283.

AR010034R