A STATISTICAL MECHANICAL APPROACH TO COMBINATORIAL CHEMISTRY

Michael W. Deem

Chemical Engineering Department, University of California, Los Angeles, California 90095-1592

I. Introduction

The goal of combinatorial chemistry is to find compositions of matter that maximize a specific material property. When combinatorial chemistry is applied to materials discovery, the desired property may be superconductivity, magnetoresistance, luminescence, ligand specificity, sensor response, or catalytic activity. When combinatorial chemistry is applied to proteins,

ADVANCES IN CHEMICAL ENGINEERING, VOL. 28 0065-2377/01 $35.00

the desired property may be enzymatic activity, fluorescence, antibiotic resistance, or substrate binding specificity. In either case, the property to be optimized, the figure of merit, is generally an unknown function of the variables and can be measured only experimentally.

Combinatorial chemistry is, then, a search over a multidimensional space of composition and noncomposition variables for regions with a high figure of merit. A traditional synthetic chemist would carry out this search by using chemical intuition to synthesize a few initial molecules. Of these molecules, those that have a favorable figure of merit would be identified. A homologous series of compounds similar to those best starting points would then be synthesized. Finally, of the compounds in these homologous series, that with the best figure of merit would be identified as the optimal material.

If the space of composition and noncomposition variables is sufficiently large, novel, or unfamiliar, the traditional synthetic approach may lead to the identification of materials that are not truly the best. It is in this case that combinatorial chemistry becomes useful. In combinatorial chemistry, trial libraries of molecules are synthesized instead of trial molecules. By synthesizing and screening for figure of merit an entire library of 10^2–10^5 molecules instead of a single molecule, the variable space can be searched much more thoroughly. In this sense, combinatorial chemistry is a natural extension of traditional chemical synthesis. Intuitive determination of the individual molecules to synthesize is replaced by methods for design of the molecular libraries. Likewise, synthesis of homologous compounds is replaced by redesign of the libraries for multiple rounds of parallel screening experiments.

While the combinatorial approach attempts to search composition space broadly, an exhaustive search is usually not possible. It would take, for example, a library of 9×10^6 compounds to search a five-component system at a mole fraction resolution of 1%. Similarly, it would take a library of $20^{100} \approx 10^{130}$ proteins to search exhaustively the space of all 100-amino acid protein domains. Clearly, a significant aspect to the design of a combinatorial chemistry experiment is the design of the library. The library members should be chosen so as to search the space of variables as effectively as possible, given the experimental constraints on the library size.

The task of searching composition space in combinatorial chemistry for regions with a high figure of merit is very similar to the task of searching configuration space by Monte Carlo computer simulation for regions with a low free energy. The space searched by Monte Carlo computer simulation is often extremely large, with 10^4 or more continuous dimensions. Yet, with recent advances in the design of Monte Carlo algorithms, one is able to locate reliably the regions with low free energy even for fairly complicated molecular systems.

This chapter pursues the analogy between combinatorial chemistry and Monte Carlo computer simulation. Examples of how to design libraries for

both materials discovery and protein molecular evolution will be given. For materials discovery, the concept of library redesign, or the use of previous experiments to guide the design of new experiments, will be introduced. For molecular evolution, examples of how to use "biased" Monte Carlo to search the protein sequence space will be given. Chemical information, whether intuition, theoretical calculations, or database statistics, can be naturally incorporated as an a priori bias in the Monte Carlo approach to library design in combinatorial chemistry. In this sense, combinatorial chemistry can be viewed as an extension of traditional chemical synthesis, one ideally suited for chemical engineering contributions.

II. Materials Discovery

A variety of materials have been optimized or developed to date by combinatorial methods. Perhaps the first experiment to gather great attention was the demonstration that inorganic oxide high-T_c superconductors could be identified by combinatorial methods (Xiang *et al.,* 1995). By searching several 128-member libraries of different inorganic oxide systems, the known compositions of superconducting BiSrCaCuO and YBaCuO were identified. Since then, many demonstrations of finding known materials and discoveries of new materials have appeared. Known compositions of giant magnetoresistant materials have been identified in libraries of various cobalt oxides (Briceño *et al.,* 1995). Blue and red phosphors have been identified from large libraries of 25,000 different inorganic oxides (Danielson *et al.,* 1988, 1997; Wang *et al.,* 1998). Polymer-based sensors for various organic vapors have been identified by combinatorial methods (Dickinson and Walt, 1997). Catalysts for the oxidation of CO to CO_2 have been identified by searching ternary compounds of Pd, Pt, and Rh or Rh, Pd, and Cu (Weinberg *et al.,* 1998; Cong *et al.,* 1999). Phase diagrams of zeolitic materials have been mapped out by a combinatorial "multiautoclave" (Akporiaye *et al.,* 1998). Novel enantioselective catalysts have been found by searching libraries of transition metal–peptide complexes (Cole *et al.,* 1996). Novel phosphatase catalysts were found by searching libraries of carboxylic acid-functionalized polyallylamine polymers (Menger *et al.,* 1995). New catalysts and conditions for C—H insertion have been found by screening of ligand–transition metal systems (Burgess *et al.,* 1996). A new catalyst for the conversion of methanol in a direct methanol fuel cell was identified by searching the quaternary compositon space of Pt, Ir, Os, and Ru (Reddington *et al.,* 1998). Finally, a novel thin-film high-dielectric compound that may be used in future generation of DRAM chips was identified by searching through over 30 multicomponent, ternary oxide systems (van Dover *et al.,* 1998).

The task of identifying the optimal compound in a materials discovery experiment can be reformulated as one of searching a multidimensional space, with the material composition, impurity levels, and synthesis conditions as variables. Present approaches to combinatorial library design and screening invariably perform a grid search in composition space, followed by a " steepest-ascent" maximization of the figure of merit. This procedure becomes inefficient in high-dimensional spaces or when the figure of merit is not a smooth function of the variables. *Indeed, the use of a grid search is what has limited essentially all current combinatorial chemistry experiments to quaternary compounds,* i.e., *to searching a space with three variables.* What is needed is an automated, yet more efficient, procedure for searching composition space.

An analogy with the computer simulation technique of Monte Carlo allows us to design just such an efficient protocol for searching the variable space (Falcioni and Deem, 2000). In materials discovery, a search is made through the composition and noncomposition variables to find good figure-of-merit values. In Monte Carlo, a search is made through configuration space to find regions of low free energy. By using insight gained from the design of Monte Carlo methods, the search in materials discovery can be improved.

A. The Space of Variables

Several variables can be manipulated to seek the material with the optimal figure of merit. Material composition is certainly a variable. But also, film thickness (van Dover *et al.,* 1998) and deposition method (Novet *et al.,* 1995) are variables for materials made in thin-film form. The processing history, such as temperature, pressure, pH, and atmospheric composition, is a variable. The guest compositon or impurity level can greatly affect the figure of merit (Cong *et al.,* 1999). In addition, the "crystallinity" of the material can affect the observed figure of merit (van Dover *et al.,* 1998). Finally, the method of nucleation or synthesis may affect the phase or morphology of the material and so affect the figure of merit (Helmkamp and Davis, 1995; Zones *et al.,* 1998).

There are important points to note about these variables. First, a small impurity composition can cause a big change in the figure of merit, as seen by the rapid variation of catalytic activity in the Cu/Rh oxidation catalyst (Cong *et al.,* 1999). Second, the phases in thin film are not necessarily the same as those in bulk, as seen in the case of the thin-film dielectric, where the optimal material was found outside the region where the bulk phase forms (van Dover *et al.,* 1998). Finally, the "crystallinity" of the material

can affect the observed figure of merit, again as seen in the thin-film dielectric example (van Dover *et al.,* 1998).

B. Library Design and Redesign

The experimental challenges in combinatorial chemistry appear to lie mainly in the screening methods and in the technology for the creation of the libraries. The theoretical challenges, on the other hand, appear to lie mainly in the library design and redesign strategies. It is this second question that is addressed by the analogy with Monte Carlo computer simulation.

Combinatorial chemistry differs from usual Monte Carlo simulations in that several simultaneous searches of the variable space are carried out. That is, in a typical combinatorial chemistry experiment, several samples, e.g., 10,000, are synthesized and screened for figure of merit at one time. With the results of this first round, a new set of samples can be synthesized and screened. This procedure can be repeated for several rounds, although current materials discovery experiments have not systematically exploited this feature.

Pursuing the analogy with Monte Carlo, each round of combinatorial chemistry corresponds to a move in a Monte Carlo simulation. Instead of tracking one system with many configurational degrees of freedom, however, many samples are tracked, each with several composition and noncomposition degrees of freedom. Modern experimental technology is what allows for the cost-effective synthesis and screening of multiple sample compositions.

The technology for materials discovery is still in the developmental stage, and future progress can still be influenced by theoretical considerations. In this spirit, I assume that the composition and noncomposition variables of each sample can be changed independently, as in spatially addressable libraries (Akporiaye *et al.,* 1998; Pirrung, 1997). This is significant, because it allows great flexibility in how the space can be searched with a limited number of experimental samples.

Current experiments uniformly tend to perform a grid search on the composition and noncomposition variables. It is preferable, however, to choose the variables statistically from the allowed values. It is also possible to consider choosing the variables in a fashion that attempts to maximize the amount of information gained from the limited number of samples screened, via a quasi-random, low-discrepancy sequence (Niederreiter, 1992; Bratley *et al.,* 1994). Such sequences attempt to eliminate the redundancy that naturally occurs when a space is searched statistically, and they have several favorable theoretical properties. An illustration of these three approaches to materials discovery library design is shown in Fig. 1.

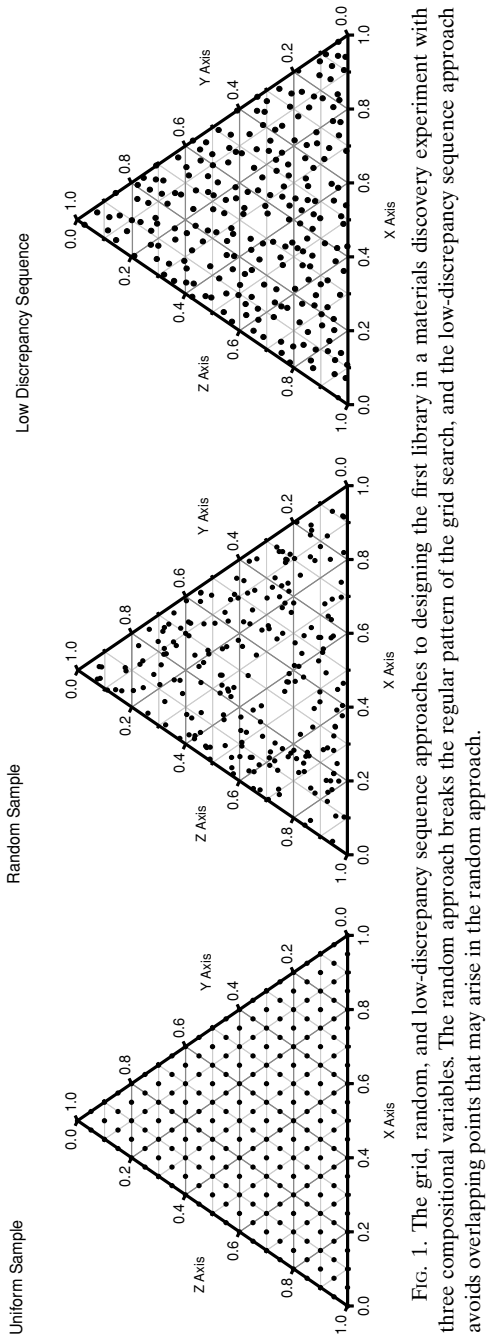

FIG. 1. The grid, random, and low-discrepancy sequence approaches to designing the first library in a materials discovery experiment with three compositional variables. The random approach breaks the regular pattern of the grid search, and the low-discrepancy sequence approach avoids overlapping points that may arise in the random approach.

Information about the figure-of-merit landscape in the composition and noncomposition variables can be incorporated by multiple rounds of screening. One convenient way to incorporate this feedback as the experiment proceeds is by treating the combinatorial chemistry experiment as a Monte Carlo in the laboratory. This approach leads to sampling the experimental figure of merit, E, proportional to exp (βE). If β is large, then the Monte Carlo procedure will seek out values of the composition and noncomposition variables that maximize the figure of merit. If β is too large, however, the Monte Carlo procedure will get stuck in relatively low-lying local maxima. The first round is initiated by choosing the composition and noncomposition variables statistically from the allowed values. The variables are changed in succeeding rounds as dictated by the Monte Carlo procedure.

Several general features of the method for changing the variables can be enumerated. The statistical method of changing the variables can be biased by concerns such as material cost, theoretical or experimental a priori insight into how the figure of merit is likely to change, and patentability. Both the composition and the noncomposition variables will be changed in each round. Likely, it would be desirable to have a range of move sizes for both types of variables. The characteristic move size would likely best be determined by fixing the acceptance ratio of the moves, as is customary in Monte Carlo simulations (Frenkel and Smit, 1996). In addition, there would likely be a smallest variable change that would be significant, due to experimental resolution limitations in the screening step. Finally, a steepest-ascent optimization to find the best local optima of the figure of merit would likely be beneficial at the end of a materials discovery experiment driven by such a Monte Carlo strategy.

C. Searching the Variable Space by Monte Carlo

Two ways of changing the variables are considered: a small random change of the variables of a randomly chosen sample and a swap of a subset of the variables between two randomly chosen samples. Swapping is useful when there is a hierarchical structure to the variables. The swapping event allows for the combination of beneficial subsets of variables between different samples. For example, a good set of composition variables might be combined with a particularly good impurity composition. Or a good set of composition variables might be combined with a good set of processing variables. These moves are repeated until all the samples in a round have been modified. The values of the figure of merit for the proposed new samples are then measured. Whether to accept the newly proposed samples or to keep the current samples for the next round is decided according to the

detailed balance acceptance criterion. For a random change of one sample, the Metropolis acceptance probability is applied

$$p_{\text{acc}}(c \to p) = \min\{1, \exp[\beta(E_{\text{proposed}} - E_{\text{current}})]\}. \tag{1}$$

Proposed samples that increase the figure of merit are always accepted; proposed samples that decrease the figure of merit are accepted with the Metropolis probability. Allowing the figure of merit occasionally to decrease is what allows samples to escape from local maxima. Moves that lead to invalid values of the composition or noncomposition variables are rejected.

For the swapping move applied to samples i and j, the modified acceptance probability is applied

$$p_{\text{acc}}(c \to p) = \min\left\{1, \exp\left[\beta\left(E^i_{\text{proposed}} + E^j_{\text{proposed}} - E^i_{\text{current}} - E^j_{\text{current}}\right)\right]\right\}. \tag{2}$$

Figure 2a shows one round of a Monte Carlo procedure. The parameter β is not related to the thermodynamic temperature of the experiment and should be optimized for best efficiency. The characteristic sizes of the random changes in the composition and noncomposition variables are also parameters that should be optimized.

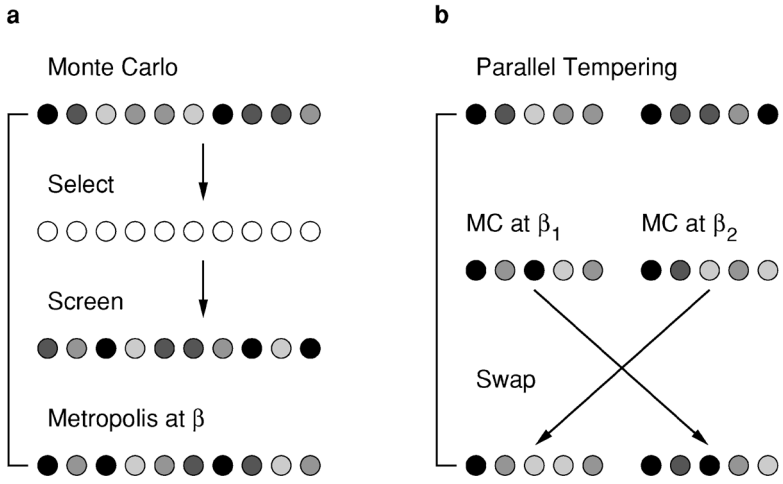

FIG. 2. Schematic of the Monte Carlo library design and redesign strategy (from Falcioni and Deem, 2000). (a) One Monte Carlo round with 10 samples: an initial set of samples, modification of the samples, measurement of the new figures of merit, and the Metropolis criterion for acceptance or rejection of the new samples. (b) One parallel tempering round with five samples at β_1 and five samples at β_2. In parallel tempering, several Monte Carlo simulations are performed at different temperatures, with the additional possibility of sample exchange between the simulations at different temperatures.

If the number of composition and noncomposition variables is too great, or if the figure of merit changes with the variables in a too-rough fashion, normal Monte Carlo will not achieve effective sampling. Parallel tempering is a natural extension of Monte Carlo that is used to study statistical (Geyer, 1991), spin glass (Marinari *et al.,* 1998), and molecular (Falcioni and Deem, 1999) systems with rugged energy landscapes. Our most powerful protocol incorporates the method of parallel tempering for changing the system variables. In parallel tempering, a fraction of the samples are updated by Monte Carlo with parameter β_1, a fraction by Monte Carlo with parameter β_2, and so on. At the end of each round, samples are randomly exchanged between the groups with different β's, as shown in Fig. 2b. The acceptance probability for exchanging two samples is

$$p_{\text{acc}}(c \rightarrow p) = \min\{1, \exp[-\Delta\beta\Delta E]\}, \tag{3}$$

where $\Delta\beta$ is the difference in the values of β between the two groups, and ΔE is the difference in the figures of merit between the two samples. It is important to notice that this exchange step does not involve any extra screening compared to Monte Carlo and is, therefore, "free" in terms of experimental costs. This step is, however, dramatically effective at facilitating the protocol to escape from local maxima. The number of different systems and the temperatures of each system are parameters that must be optimized.

To summarize, the first round of combinatorial chemistry consists of the following steps: constructing the initial library of samples, measuring the initial figures of merit, changing the variables of each sample a small random amount or swapping subsets of the variables between pairs of samples, constructing the proposed new library of samples, measuring the figures of merit of the proposed new samples, accepting or rejecting each of the proposed new samples, and performing parallel tempering exchanges. Subsequent rounds of combinatorial chemistry repeat these steps, starting with making changes to the current values of the composition and noncomposition variables. These steps are repeated for as many rounds as desired, or until maximal figures of merit are found.

D. The Simplex of Allowed Compositions

The points to be sampled in materials discovery are the allowed values of the composition and noncomposition variables. Typically, the composition variables are specified by the mole fractions. Since the mole fractions sum to one, sampling on these variables requires special care.

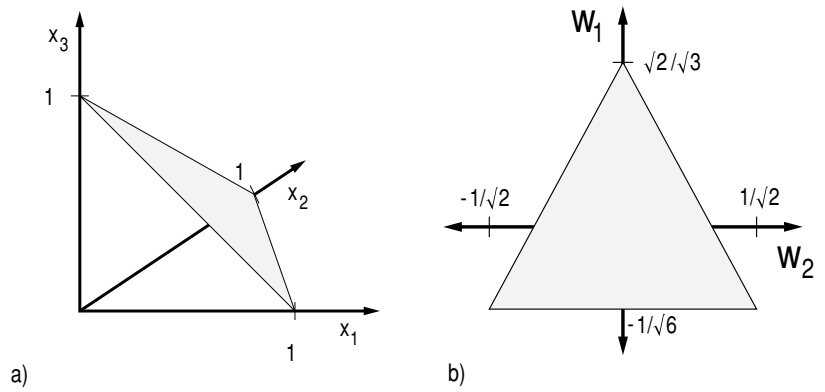

FIG. 3. The allowed composition range of a three-component system is shown in (a) the original composition variables, x_i, and (b) the Gram–Schmidt variables, w_i.

In particular, the specification or modification of the d mole fraction variables, x_i, is done in the $(d-1)$-dimensional hyperplane orthogonal to the d-dimensional vector $(1, 1, \ldots, 1)$. This procedure ensures that the constraint $\sum_{i=1}^{d} x_i = 1$ is maintained. This subspace is identified by a Gram–Schmidt procedure, which identifies a new set of basis vectors, $\{\mathbf{u}_i\}$, that span this hyperplane. Figure 3 illustrates the geometry for the case of three composition variables.

The new basis set is identified as follows. First, $\mathbf{u}_d$ is defined to be the unit vector orthogonal to the allowed hyperplane

$$\mathbf{u}_d = \left(\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}, \ldots, \frac{1}{\sqrt{d}}\right). \tag{4}$$

The remaining $\mathbf{u}_i$, $1 \leq i < d$, are chosen to be orthogonal to $\mathbf{u}_d$, so that they lie in the allowed hyperplane. Indeed, the $\mathbf{u}_i$ form an orthonormal basis for the composition space. This orthonormal basis is identified by the Gram–Schmidt procedure. First, the original composition basis vectors are defined

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, \ldots, 0, 0) \\ \mathbf{e}_2 &= (0, 1, \ldots, 0, 0) \\ &\vdots \\ \mathbf{e}_{d-1} &= (0, 0, \ldots, 1, 0). \end{aligned} \tag{5}$$

Each $\mathbf{u}_i$ is identified by projecting these basis vectors onto the space

orthogonal to $\mathbf{u}_d$ and the $\mathbf{u}_j$, $j < i$, already identified

$$\begin{aligned}
\mathbf{u}_1 &= \frac{\mathbf{e}_1 - (\mathbf{e}_1 \cdot \mathbf{u}_d)\mathbf{u}_d}{|\mathbf{e}_1 - (\mathbf{e}_1 \cdot \mathbf{u}_d)\mathbf{u}_d|} \\
\mathbf{u}_2 &= \frac{\mathbf{e}_2 - (\mathbf{e}_2 \cdot \mathbf{u}_d)\mathbf{u}_d - (\mathbf{e}_2 \cdot \mathbf{u}_1)\mathbf{u}_1}{|\mathbf{e}_2 - (\mathbf{e}_2 \cdot \mathbf{u}_d)\mathbf{u}_d - (\mathbf{e}_2 \cdot \mathbf{u}_1)\mathbf{u}_1|} \\
&\vdots \\
\mathbf{u}_i &= \frac{\mathbf{e}_i - (\mathbf{e}_i \cdot \mathbf{u}_d)\mathbf{u}_d - \sum_{j=1}^{i-1}(\mathbf{e}_i \cdot \mathbf{u}_j)\mathbf{u}_j}{|\mathbf{e}_i - (\mathbf{e}_i \cdot \mathbf{u}_d)\mathbf{u}_d - \sum_{j=1}^{i-1}(\mathbf{e}_i \cdot \mathbf{u}_j)\mathbf{u}_j|}.
\end{aligned} \tag{6}$$

A point in the allowed composition range is specified by the vector $\mathbf{x} = \sum_{i=1}^{d} w_i \mathbf{u}_i$, with $w_d = 1/\sqrt{d}$. Note that the values w_i are related to the composition values x_i by a rotation matrix, since the Gram–Schmidt procedure simply identifies a rotated basis for the composition space

$$\mathbf{x} = R\mathbf{w}, \tag{7}$$

where R_{ij} is given by the ith component of $\mathbf{u}_j$. Each of the numbers w_i, $1 \leq i < d$, is to be varied in the materials discovery experiment. Not all values of w_i are feasible, however, since the constraint $x_i \geq 0$ must be satisfied. Feasible values are identified by transforming the w_i to the x_i by Eq. (7), and then checking that the compositon variables are nonnegative. The constraint that the composition variables sum to unity is automatically ensured by the choice $w_d = 1/\sqrt{d}$.

E. Significance of Sampling

Sampling the figure of merit by Monte Carlo, rather than global optimization by some other method, is favorable for several reasons. First, Monte Carlo is an effective stochastic optimization method. Second, simple global optimization may be misleading since concerns such as patentability, cost of materials, and ease of synthesis are not usually included in the experimental figure of merit. Moreover, the screen that is most easily performed in the laboratory, the "primary screen," is usually only roughly correlated with the true figure of merit. Indeed, after finding materials that look promising based upon the primary screen, experimental secondary and tertiary screens are usually performed to identify that material which is truly optimal. Third, it might be advantageous to screen for several figures of merit at once. For example, it might be profitable to search for reactants and conditions that lead to the synthesis of several zeolites with a particularly favorable property, such as the presence of a large pore. As another example, it might be useful

to search for several electrocatalysts that all possess a useful property, such as being able to serve as the anode or cathode material in a particular fuel cell.

For all of these reasons, sampling by Monte Carlo to produce several candidate materials is preferred over global optimization.

F. The Random Phase Volume Model

The ultimate test of new, theoretically motivated protocols for materials discovery is, of course, experimental. To motivate such experimentation, the effectiveness of these protocols is demonstrated by combinatorial chemistry experiments where the experimental screening step is replaced by figures of merit returned by the random-phase volume model. The random phase volume model is not fundamental to the protocols; it is introduced as a simple way to test, parameterize, and validate the various searching methods.

The random phase volume model relates the figure of merit to the composition and noncomposition variables in a statistical way. The model is fast enough to allow for validation of the proposed searching methods on an enormous number of samples yet possesses the correct statistics for the figure-of-merit landscape.

The composition mole fractions are nonnegative and sum to unity, and so the allowed compositions are constrained to lie within a simplex in $d-1$ dimensions. For the familiar ternary system, this simplex is an equilateral triangle, as shown in Fig. 3b. Typically, several phases will exist for different compositions of the material. The figures of merit will be dramatically different between each of these distinct phases. To mimic this expected behavior, the composition variables are grouped in the random phase volume model into phases centered around N_x points $\mathbf{x}_\alpha$ randomly placed within the allowed composition range. The phases form a Voronoi diagram (Sedgewick, 1988), as shown in Fig. 4.

The random phase volume model is defined for any number of composition variables, and the number of phase points is defined by requiring the average spacing between phase points to be $\xi = 0.25$. To avoid edge effects, additional points are added in a belt of width 2ξ around the simplex of allowed compositions. The number of phase points for different grid spacing is shown in Table I.

The figure of merit should change dramatically between composition phases. Moreover, within each phase α, the figure of merit should also vary with $\mathbf{y} = \mathbf{x} - \mathbf{x}_\alpha$ due to crystallinity effects such as crystallite size, intergrowths, defects, and faulting (van Dover *et al.,* 1998). In addition, the noncomposition variables should also affect the measured figure of merit. The noncomposition variables are denoted by the b-dimensional vector **z,** with

TABLE I
NUMBER OF PHASE POINTS AS A FUNCTION OF DIMENSION AND SPACING

ξ	d	Number of points
0.1	3	193
0.1	4	1,607
0.1	5	12,178
0.1	6	81,636
0.2	3	86
0.2	4	562
0.2	5	3,572
0.2	6	20,984
0.25	3	70
0.25	4	430
0.25	5	2,693
0.25	6	15,345
0.3	3	59
0.3	4	353
0.3	5	2,163
0.3	6	12,068
0.35	3	53
0.35	4	306
0.35	5	1,850
0.35	6	10,234

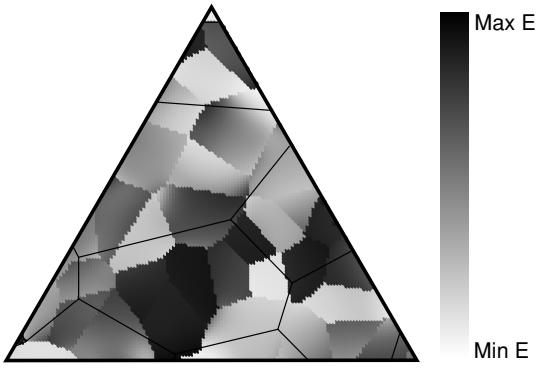

FIG. 4. The random phase volume model (from Falcioni and Deem, 2000). The model is shown for the case of three composition variables and one noncomposition variable. The boundaries of the **x** phases are evident by the sharp discontinuities in the figure of merit. To generate this figure, the **z** variable was held constant. The boundaries of the **z** phases are shown as thin dark lines.

each component constrained to fall within the range $[-1, 1]$ without loss of generality. There can be any number of noncomposition variables. The figure of merit depends on the composition and noncomposition variables in a correlated fashion. In particular, how the figure of merit changes with the noncomposition variables should depend on the values of the compositon variables. To mimic this behavior within the random phase volume model, the noncomposition variables also fall within N_z noncomposition phases defined in the space of composition variables. There are a factor of 10 fewer noncomposition phases than composition phases.

The functional form of the model when $\mathbf{x}$ is in composition phase α and noncomposition phase γ is

$$E(\mathbf{x}, \mathbf{z}) = U_\alpha + \sigma_x \sum_{k=1}^{q} \sum_{i_1 \geq \ldots \geq i_k = 1}^{d} f_{i_1 \ldots i_k} \xi_x^{-k} A_{i_1 \ldots i_k}^{(\alpha k)} y_{i_1} y_{i_2} \ldots y_{i_k} + \frac{1}{2} \left(W_\gamma + \sigma_z \sum_{k=1}^{q} \sum_{i_1 \geq \ldots \geq i_k = 1}^{b} f_{i_1 \ldots i_k} \xi_z^{-k} B_{i_1 \ldots i_k}^{(\gamma k)} z_{i_1} z_{i_2} \ldots z_{i_k} \right), \tag{8}$$

where $f_{i_1 \ldots i_k}$ is a constant symmetry factor, ξ_x and ξ_z are constant scale factors, and U_α, W_γ, $A_{i_1 \ldots i_k}^{(\alpha k)}$, and $B_{i_1 \ldots i_k}^{(\gamma k)}$ are random Gaussian variables with unit variance. In more detail, the symmetry factor is given by

$$f_{i_1 \ldots i_k} = \frac{k!}{\prod_{i=1}^{l} o_i!}, \tag{9}$$

where l is the number of distinct integer values in the set $\{i_1, \ldots, i_k\}$, and o_i is the number of times that distinct value i is repeated in the set. Note that $1 \leq l \leq k$ and $\sum_{i=1}^{l} o_i = k$. The scale factors are chosen so that each term in the multinomial contributes roughly the same amount: $\xi_x = \xi/2$ and $\xi_z = (\langle z^6 \rangle / \langle z^2 \rangle)^{1/4} = (3/7)^{1/4}$. The σ_x and σ_z are chosen so that the multinomial, crystallinity terms contribute 40% as much as the constant, phase terms on average. For both multinomials $q = 6$. As Fig. 4 shows, the random phase volume model describes a rugged figure-of-merit landscape, with subtle variations, local maxima, and discontinuous boundaries.

G. Several Monte Carlo Protocols

Six ways of searching the variable space are tested with increasing numbers of composition and noncomposition variables. The total number of samples whose figure of merit will be measured is fixed at $M = 100,000$, so that all protocols have the same experimental cost. The single-pass protocols grid, random, and low-discrepancy sequence (LDS) are considered.

For the grid method, the number of samples in the composition space is $M_x = M^{(d-1)/(d-1+b)}$ and the number of samples in the noncomposition space is $M_z = M^{b/(d-1+b)}$. The grid spacing of the composition variables is $\zeta_x = (V_d/M_x)^{1/(d-1)}$, where

$$V_d = \frac{\sqrt{d}}{(d-1)!} \tag{10}$$

is the volume of the allowed composition simplex. Note that the distance from the centroid of the simplex to the closest point on the boundary of the simplex is

$$R_d = \frac{1}{[d(d-1)]^{1/2}}. \tag{11}$$

The spacing for each component of the noncomposition variables is $\zeta_z = 2/M_z^{1/b}$. For the LDS method, different quasi-random sequences are used for the composition and noncomposition variables. The feedback protocols Monte Carlo, Monte Carlo with swap, and parallel tempering are considered. The Monte Carlo parameters were optimized on test cases. It was optimal to perform 100 rounds of 1000 samples with $\beta = 2$ for $d = 3$ and $\beta = 1$ for $d = 4$ or 5, and $\Delta x = 0.1R_d$ and $\Delta z = 0.12$ for the maximum random displacement in each component. The swapping move consisted of an attempt to swap all of the noncomposition values between the two chosen samples, and it was optimal to use $P_{\text{swap}} \simeq 0.1$ for the probability of a swap versus a regular random displacement. For parallel tempering it was optimal to perform 100 rounds with 1000 samples, divided into three subsets: 50 samples at $\beta_1 = 50$, 500 samples at $\beta_2 = 10$, and 450 samples at $\beta_3 = 1$. The 50 samples at large β essentially perform a "steepest-ascent" optimization and have smaller $\Delta x = 0.01R_d$ and $\Delta z = 0.012$.

H. Effectiveness of the Monte Carlo Strategies

The figures of merit found by the protocols are shown in Fig. 5. The single-round protocols, random and low-discrepancy sequence, find better solutions than does grid in one round of experiment. Interestingly, the low-discrepancy sequence approach fares no better than does random, despite the desirable theoretical properties of low-discrepancy sequences.

The multiple-round, Monte Carlo protocols appear to be especially effective on the more difficult systems with larger numbers of composition and noncomposition variables. That is, the Monte Carlo methods have a tremendous advantage over one-pass methods, especially as the number of variables increases, with parallel tempering the best method. The Monte Carlo

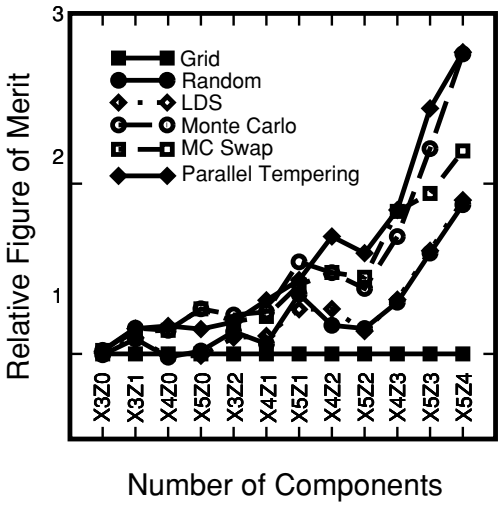

FIG. 5. The maximum figure of merit found with different protocols on systems with different number of composition (**x**) and noncomposition (**z**) variables (from Falcioni and Deem, 2000). The results are scaled to the maximum found by the grid searching method. Each value is averaged over scaled results on 10 instances of the random phase volume model with different random phases. The Monte Carlo methods are especially effective on systems with a larger number of variables, where the maximal figures of merit are more difficult to locate.

methods, in essence, gather more information about how best to search the variable space with each succeeding round. This feedback mechanism proves to be effective even for the relatively small total sample size of 100,000 considered here. It is expected that the advantage of the Monte Carlo methods will become even greater for larger sample sizes. Note that in cases such as catalytic activity, sensor response, or ligand specificity, the experimental figure of merit would likely be exponential in the values shown in Fig. 5, so that the success of the Monte Carlo methods would be even more dramatic. A better calibration of the parameters in Eq. (8) may be possible as more data become available in the literature.

I. ASPECTS OF FURTHER DEVELOPMENT

The space of composition and noncomposition variables to search in materials discovery experiments can be forbiddingly large. Yet, by using Monte Carlo methods, one can achieve an effective search with a limited number of experimental samples.

Efficient implementations of the Monte Carlo search strategies are feasible with existing library creation technology. Moreover "closing the loop"

between library design and redesign is achievable with the same database technology currently used to track and record the data from combinatorial chemistry experiments. These multiple-round protocols, when combined with appropriate robotic automation, should allow the practical application of combinatorial chemistry to more complex and interesting systems.

Many details need to be worked out to flesh out the proposed protocols for materials discovery. For example,

1. How rough are real figures of merit, and can the random phase volume model be calibrated better?
2. Can more of the hierarchical structure of the variables be identified? and
3. What are the best methods of manipulating the variables in the Monte Carlo?

Additional questions, some of which this chapter has begun to answer, include how does the proximity to the global optimum scale with the number of samples and with the algorithm by which they are selected? What is the best set of samples to choose for an optimal result, chosen "all at once" or in stages or sequentially? What is the minimum number of samples required to make a Monte Carlo-based algorithm attractive as the driver?

III. Protein Molecular Evolution

The space to be searched in protein combinatorial chemistry experiments is extremely large. Consider, for example, that a relatively short 100-amino acid protein domain were to be evolved. The number of possible amino acid sequences of this length is $20^{100} \approx 10^{130}$, since there are 20 naturally occurring amino acid residues. Clearly, all of these sequences cannot be synthesized and then screened for figure of merit in the laboratory. Some means must be found for searching this space with the 10^4 or so proteins that can be screened per day experimentally.

A hierarchical decomposition of the protein space can provide an effective searching procedure. It is known from protein structural biology that proteins are encoded by DNA sequences, DNA sequences code for amino acids, amino acids arrange into secondary structures, secondary structures arrange into domains, domains group to form protein monomers, and protein monomers aggregate to form multiprotein complexes. By sampling on each level of this hierarchy, one is able to search the sequence space much more effectively. In this chapter, search strategies making use of the DNA, amino acid, and secondary structure hierarchy will be described. With this

approach, functional protein space has a large, yet manageable, number of dimensions. That is, in a 100-amino acid protein domain there are approximate 10 secondary structures of 5 types (helices, loops, strands, turns, and others), roughly yielding the potential for $\approx 10^7$ basic protein folds. Organization into secondary structural classes represents a dramatic reduction in the complexity of sequence space, since there are $\approx 10^{170}$ different DNA sequences and $\approx 10^{130}$ different amino acid sequences in this space.

Sampling on the different levels of protein structure is analogous to combination of different move types in a Monte Carlo simulation. A variety of moves, from small, local moves to large, global moves, are often incorporated in the most successful Monte Carlo simulations. While protein molecular evolution is carried out in the laboratory, and Monte Carlo simulations are carried out *in silico,* the parallels are striking. One of the most powerful new concepts in Monte Carlo is the idea that moves should be "biased" (Frenkel and Smit, 1996). That is, small moves, such as the Metropolis method, sample configuration space rather slowly. Larger moves are preferred, since they sample the space more rapidly. Large moves are usually rejected, however, since they often lead the system into a region of high energy. So that the large moves will be more successful, a bias toward regions that look promising is included. Such biased Monte Carlo simulations have been a factor of 10^5 to 10^{10} times more efficient than previous methods, and they have allowed the examination of systems previously uncharacterizable by molecular simulation techniques (Frenkel *et al.,* 1992; Frenkel and Smit, 1992; de Pablo *et al.,* 1992; Smit and Maesen, 1995).

In this section, the possibility of evolving protein molecules by strategies similar to biased Monte Carlo will be explored. The large moves of Monte Carlo are implemented by changing an evolving protein at the secondary structure level. These evolutionary events will be biased, in that the amino acid sequences inserted will be chosen so that they code for viable secondary structures. The concept of bias also applies at the amino acid level, where different DNA sequences coding for the same amino acids can lead to different propensities for future evolution.

A. What Is Protein Molecular Evolution?

Protein molecular evolution can be viewed as combinatorial chemistry of proteins. Since protein sequence space is so large, most experiments to date have sought to search only small regions. A typical experiment seeks to optimize the figure of merit of an existing protein. For example, an improvement in the selectivity or activity of an enzyme might be sought. Alternatively, an expansion in the operating range of an enzyme might be sought to higher

temperatures or pressures. This improvement would be achieved by changing, or evolving, the amino acid sequence of the enzyme.

A more ambitious goal would be the *ab initio* evolution of a protein with a specific function. That is, nothing would be known about the desired molecule, except that it should be a protein and that an experimental screen for the desired figure of merit is available. One might want, for example, an enzyme that catalyzes an unusual reaction. Or one might want a protein that binds a specific substrate. Or one might want a protein with an unusual fluorescence spectrum. The *ab initio* evolution of a protein has never been accomplished before. Such a feat would be remarkable. Natural biological diversity has evolved despite the essentially infinite complexity of protein sequence. Replication of this feat in the laboratory would represent substantial progress, and mimicking this feat of Nature is a current goal in the molecular evolution field. Indeed, the protocols described in this section are crafted with just this task in mind.

A still more ambitious goal would be the evolution of a multiprotein complex. This is a rather challenging task, due to the increased complexity of the space to be searched. The task can be made manageable by asking a rather general evolutionary question. One can seek to evolve, for example, a multiprotein complex that can serve as the coat protein complex for a virus. Since there are many proteins that may accomplish this task, this evolutionary task may not be as specific and difficult as it might seem at first.

The most ambitious goal for laboratory evolution that has been imagined is the evolution of new life forms. Evolution on this scale requires changes not only at the secondary structure scale, but also at the domain, protein, and protein pathway scale. Due to their simplicity, viruses or phage would be the most likely targets of such large-scale evolution attempts. It is unclear how new life forms would be distributed in terms of pathogenicity, and so such experiments should be approached with caution.

The hierarchical decomposition of sequence space will allow effective molecular evolution if there are many proteins with a high value of any particular figure of merit. That is, if only 1 of 10^{130} small protein domains exhibits a high score on a particular figure of merit, this protein is unlikely to be identified. On the other hand, if many proteins score highly on the figure of merit, only a subset of these molecules need be sampled. This same issue arises in conventional Monte Carlo simulations. Sampling all of configuration space is never possible in a simulation, yet ensemble averages and experimental behavior can be reproduced by sampling representative configurations. That life on our planet has evolved suggests that there is a great redundancy in protein space (Kauffman, 1993), and so one may hope to search this space experimentally with sufficiently powerful moves.

B. Background on Experimental Molecular Evolution

There are some constraints on molecular evolution as it is carried out in the laboratory. There are constraints arising from limitations of molecular biology, i.e., only certain types of moves are possible on the DNA that codes for the protein. There are also constraints arising from technical limitations, i.e., only a certain number of proteins can be screened for figure of merit in a day.

Existing approaches to the evolution of general proteins are essentially limited to changes at the single-base level. Somewhat more sophisticated methods are available for evolution of antibodies, but this is a special case that is not considered here. The first type of evolutionary change that is possible in the laboratory is a base substitution. Base substitutions are naturally made as DNA is copied or amplified by PCR. The rate at which base substitutions, or mistakes in the copying of the template DNA, are made can be adjusted by varying experimental conditions, such as the manganese and magnesium ion concentrations. It is important to note that these base substitutions are made without knowledge of the DNA sequences of the evolving proteins. Equally important is that these base substitutions are made without the use of a chemical synthesizer. These changes are made naturally within the context of efficient molecular biology methods. Another means of modifying an existing protein is to use random or directed mutagenesis to change specific DNA bases so that they code for random or specified amino acids. The approach requires both that the DNA sequence be known and that it subsequently be synthesized by chemical means. Such a laborious approach is not practical in high-throughput evolution experiments, where typically 10^4 proteins are simultaneously modified and evolved per day.

Since the average length of a human gene is roughly 1800 bases, base-by-base point mutation will achieve significant evolution only very slowly. More significantly, the figure-of-merit landscape for protein function is typically quite rugged. Base mutation, therefore, invariably ceases evolution at a local optima of the figure of merit. Base mutation can be viewed as an experimental method for local optimization of protein figures of merit.

Much of the current enthusiasm for protein molecular evolution is due to the discovery of DNA shuffling by Pim Stemmer in 1994. DNA shuffling is a method for evolving an existing protein to achieve a higher figure of merit. The great genius of Stemmer was to develop a method for combining beneficial base mutations that is naturally accomplished with the tools of molecular biology and that does not require DNA sequencing or chemical synthesis. The method is successful because combination of base mutations that were individually beneficial is likely to lead to an even higher figure of merit than is achieved by either mutation alone. Of course, this will not

TABLE II
GENES AND OPERONS EVOLVED BY DNA SHUFFLING (FROM PATTEN *ET AL.*, 1997)[a]

System	Improvement	Size	Mutations
TEM-1 β-lactamase	Enzyme activity 32,000-fold	333 aa	6 aa
β-Galactosidase	Fucosidase activity 66-fold	1333 aa	6 aa
Green fluorescence protein	Protein folding 45-fold	266 aa	3 aa
Antibody	Avidity >400-fold	233 aa	34 aa[b]
Antibody	Expression level 100-fold	233 aa	5 aa
Arsenate operon	Arsenate resistance 40-fold	766 aa	3 aa
Alkyl transferase	DNA repair 10-fold	166 aa	7 aa
Benzyl esterase	Antibiotic deprotection 150-fold	500 aa	8 aa
tRNA synthetase	Charging of engineered tRNA 180-fold	666 aa	ND

[a] aa, amino acids; ND, not determined.
[b] This was a case of family shuffling (Crameri *et al.,* 1998), so most of these changes were between homologous amino acids.

always be true, but the extent to which it is true is the extent to which DNA shuffling will be an effective technique. DNA shuffling, combined with base mutation, is the current state of the art experimental technique for protein molecular evolution.

Table II lists a few of the protein systems that have been evolved by the Stemmer group. Evidently, DNA shuffling is highly effective at improving the function of an existing protein, much more effective than is simple base mutation. The specificity of an enzyme can even be altered, as in the conversion of a β-galactosidase into a fucosidase. A rough median of the improvement factors is about 100. Most important, however, is that all of these improvements were achieved with a relatively small number of amino acid changes. On average, only 6 amino acids were altered, of roughly 400 total residues in the protein. As with base mutation, then, DNA shuffling is able to search sequence space only locally. After a small number of amino acid changes, DNA shuffling produces a protein with a locally rather than globally optimal figure of merit.

The current state-of-the-art experimental techniques for protein molecular evolution can be viewed as local optimization procedures in protein

sequence space. Alternatively, they can be viewed as experimental implementations of a simple, or Metropolis, Monte Carlo procedure. By using our intuition regarding the design of powerful, biased Monte Carlo algorithms, we can develop more powerful experimental protocols for molecular evolution.

Interestingly, theoretical treatments of evolution, whether in Nature or in the laboratory, tend to consider only the effects of point mutation (Kauffman, 1993; Volkenstein, 1994). Indeed, interesting theories regarding the evolutionary potential of point mutations have been developed. As shown experimentally, however, point mutation is incapable of significantly evolving proteins at substantial rate. Even the more powerful technique of DNA shuffling searches protein space merely locally. Only with the inclusion of more dramatic moves, such as changes at the level of secondary structures, can protein space be searched more thoroughly.

C. The Generalized NK Model

To validate the molecular evolution protocols to be presented, a model that relates amino acid sequence to protein function is needed. Of course, the real test of these protocols should be experimental, and I hope that these experiments will be forthcoming. To stimulate interest in the proposed protocols, their effectiveness will be simulated on a model of protein function. Such a model would seem to be difficult to construct. It is extremely difficult to determine the three-dimensional structure of a protein given the amino acid sequence. Moreover, it is extremely difficult to calculate any of the typical figures of merit given the three-dimensional structure of a protein.

It is fortunate that a model that relates figure of merit to amino acid sequence for a specific protein is not needed. The requirement is simply a model that produces figure-of-merit landscapes in sequence space that are analogous to those that would be measured in the laboratory on an ensemble of proteins. This type of model is easier to construct, and a random energy model can be used to accomplish the task.

The generalized NK model is just such a random energy model. The NK model was first introduced to model combinatorial chemistry experiments on peptides (Kauffman and Levin, 1987; Kauffman, 1993; Kauffman and MacReady, 1995). It was subsequently generalized to account for secondary structure in real proteins (Perelson and Macken, 1995). The model was further generalized to account for interactions between the secondary structures and for the presence of a binding pocket (Bogarad and Deem, 1999).

This generalized NK model assigns a unique figure of merit to each evolving protein sequence. This model, while a simplified description of real

proteins, captures much of the thermodynamics of protein folding and ligand binding. The model takes into account the formation of secondary structures via the interactions of amino acid side chains as well as the interactions between secondary structures within proteins. In addition, for specificity, the figure of merit is assumed to be a binding constant, and so the model includes a contribution representing binding to a substrate. The combined ability to fold and bind substrate is what will be optimized or evolved. That is, the direction of the protein evolution will be based upon the figure of merit returned by this generalized NK model. This generalized NK model contains several parameters, and a reasonable determination of these parameters is what allows the model to compare successfully with experiment.

The specific energy function used as the selection criterion in the molecular simulations is

$$U = \sum_{\alpha=1}^{M} U_{\alpha}^{\mathrm{sd}} + \sum_{\alpha>\gamma=1}^{M} U_{\alpha\gamma}^{\mathrm{sd-sd}} + \sum_{i=1}^{P} U_{i}^{\mathrm{c}}. \tag{12}$$

This energy function is composed of three parts: secondary structural subdomain energies (U^{sd}), subdomain–subdomain interaction energies ($U^{\mathrm{sd-sd}}$), and chemical binding energies (U^{c}). Each of these three energy terms is weighted equally, and each has a magnitude near unity for a random sequence of amino acids. In this NK-based simulation, each different type of amino acid behaves as a completely different chemical entity; therefore, only $Q = 5$ chemically distinct amino classes are considered (e.g., negative, positive, polar, hydrophobic, and other). Interestingly, restricted alphabets of amino acids not only are capable of producing functional proteins (Kamtekar *et al.,* 1993; Riddle *et al.,* 1997) but also may have been used in the primitive genetic code (Miller and Orgel, 1974; Schuster and Stadler, 1998). The evolving protein will be a relatively short, 100-amino acid protein domain. Within this domain will be roughly $M = 10$ secondary structural subdomains, each $N = 10$ amino acids in length. The subdomains belong to one of $L = 5$ different types (*e.g.,* helices, strands, loops, turns, and others). This gives L different (U^{sd}) energy functions of the NK form (Kauffman and Levin, 1987; Kauffman, 1993; Kauffman and MacReady, 1995; Perelson and Macken, 1995),

$$U_{\alpha}^{\mathrm{sd}} = \frac{1}{[M(N-K)]^{1/2}} \sum_{j=1}^{N-K+1} \sigma_{\alpha}(a_j, a_{j+1}, \ldots, a_{j+K-1}). \tag{13}$$

The degree of complexity in the interactions between the amino acids is parameterized by the value of K. Low values of K lead to figure-of-merit landscapes upon which evolution is easy, and high values of K lead to extremely rugged landscapes upon which evolution is difficult. Combinatorial

chemistry experiments on peptides have suggested the value of $K = 4$ as a reasonable one (Kauffman and MacReady, 1995). Note that the definition of K here is one greater than the convention in (Kauffman and Levin, 1987; Kauffman, 1993; Kauffman and MacReady, 1995). The quenched, unit-normal random number σ_α in Eq. (13) is different for each value of its argument for each of the L classes. This random form mimics the complicated amino acid side-chain interactions within a given secondary structure. The energy of interaction between secondary structures is given by

$$U_{\alpha\gamma}^{\text{sd}-\text{sd}} = \left[\frac{2}{DM(M-1)}\right]^{1/2} \times \sum_{i=1}^{D} \sigma_{\alpha\gamma}^{(i)} \left(a_{j_1}^{\alpha}, \ldots, a_{j_{K/2}}^{\alpha}; a_{j_{K/2+1}}^{\gamma}, \ldots, a_{j_K}^{\gamma}\right). \quad (14)$$

The number of interactions between secondary structures is set at $D = 6$. Here the unit-normal weight, $\sigma_{\alpha\gamma}^{(i)}$, and the interacting amino acids, $\{j_1, \ldots, j_K\}$, are selected at random for each interaction (i, α, γ). The chemical binding energy of each amino acid is given by

$$U_i^{\text{c}} = \frac{1}{\sqrt{P}} \sigma_i(a_i). \quad (15)$$

The contributing amino acid, i, and the unit-normal weight of the binding, σ_i, are chosen at random. A typical binding pocket is composed of five amino acids, and so the choice of $P = 5$ is made.

D. Experimental Conditions and Constraints

A typical protein evolution experiment starts with an initial protein sequence. This sequence is then copied to a large number of identical sequences. All of these sequences are evolved, or mutated, in parallel. After one round of mutation events, the proteins are screened for figure of merit. This screening step is typically the rate-limiting step, and so the efficiency of this step determines how many proteins can be evolved in parallel. For typical figures of merit, 10,000 proteins can be screened in a day. If selection, that is, use of a screen based upon whether an organism lives or dies, were performed instead, 10^9–10^{15} proteins could be screened in a day. Selection is a special case, however, and so the more conservative case of screening 10,000 proteins per day is considered.

After the screen, the proteins are ranked according to their measured value of the figure of merit. Typically, the top x percent of the sequences is kept for the next round of mutation. The parameter x is to be adjusted experimentally. In the simulated evolutions, the value of $x = 10\%$ was always found to be optimal. Other methods for selecting the proteins to keep for

the next round have been considered. For example, keeping proteins proportional to exp $(-\beta U)$ has been considered. This strategy seems to work less well than the top x percent method. The main reason seems to be that in the top x percent method, the criterion for selecting which sequences to keep adjusts naturally with the range of figures of merit found in the evolving sequences. After the top x percent sequences are selected, they are copied back up to a total of 10,000 sequences. These sequences are the input for the next round of mutation and selection.

In the simulated molecular evolutions, the experiment is continued for 100 rounds. This is a relatively large number of rounds to carry out experimentally. With the most powerful protocols, however, it is possible to evolve proteins *ab initio*. This feat has not been achieved to date in the laboratory. To mimic this feat of Nature, one should be willing to do some number of rounds.

E. Several Hierarchical Evolution Protocols

1. Amino Acid Substitution

To obtain a baseline for searching fold space, molecular evolution is first simulated via simple mutagenesis (see Fig. 6a). Simulated evolutions by amino acid substitution lead to significantly improved protein energies, as shown in Table III. These evolutions always terminated at local energy minima, however. This trapping is due to the difficulty of combining the large number of correlated substitutions necessary to generate new protein folds. Increasing the screening stringency in later rounds did not improve the binding constants of simulated proteins, most likely due to the lack of additional selection criteria such as growth rates. Although only nonconservative mutations were directly simulated, conservative and synonymous neutral mutations are not excluded and can be taken into account in a more detailed treatment. Indeed, the optimized average mutation rate of 1 amino acid substitution/sequence/round is equivalent to roughly 1–6 random base substitutions/round.

2. DNA Shuffling

DNA shuffling improves the search of local fold space via a random yet correlated combination of homologous coding fragments that contain limited numbers of beneficial amino acid substitutions. As in experimental evolutions (Stemmer, 1994; Crameri *et al.,* 1998; Zhang *et al.,* 1997; Moore *et al.,* 1997), the simulated shuffling improved protein function significantly better than did point mutation alone (see Table III and Fig. 6b). However, local

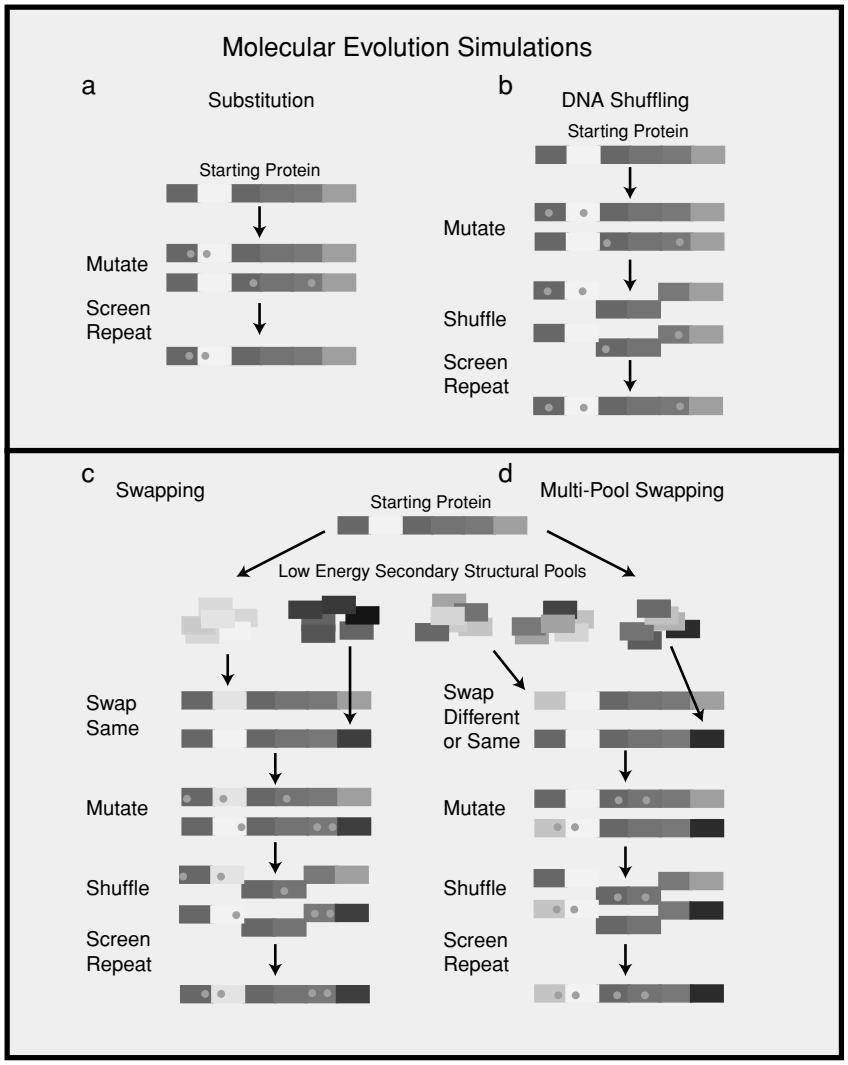

FIG. 6. Schematic diagram of the simulated molecular evolution protocols (from Bogarad and Deem, 1999). (a) Simulation of molecular evolution via base substitution (substitutions are represented by orange dots). (b) Simulated DNA shuffling showing the optimal fragmentation length of two subdomains. (c) The hierarchical optimization of local space searching: The 250 different sequences in each of the five pools (e.g., helices, strands, turns, loops, and others) are schematically represented by different shades of the same color. (d) The multipool swapping model for searching vast regions of tertiary fold space is essentially the same as in c except that now sequences from all five structural pools can be swapped into any subdomain. Multipool swapping allows for the formation of new tertiary structures by changing the type of secondary structure at any position along the protein.

TABLE III
RESULTS OF MONTE CARLO SIMULATION OF THE EVOLUTION PROTOCOLS (FROM BOGARAD AND DEEM, 1999)[a]

Evolution method	U_{start}	$U_{evolved}$	$k_{binding}$
Amino acid substitution	−17.00	−23.18	1
DNA shuffling	−17.00	−23.83	100
Swapping	0	−24.52	1.47×10^4
Mixing	0	−24.88	1.81×10^5
Multipool swapping[b]	0	−25.40	8.80×10^6

[a] The starting polypeptide energy of −17.00 comes from a protein-like sequence (minimized U^{sd}), and 0 comes from a random initial sequence of amino acids. The evolved energies and binding constants are median values. The binding constants are calculated from $k_{binding} = ae^{-bU}$, where a and b are constants determined by normalizing the binding constants achieved by point mutation and shuffling to 1 and 100, respectively.

[b] Note that the energies and binding constants achieved via multipool swapping represent typical best-evolved protein folds.

barriers in the energy function also limit molecular evolution via DNA shuffling. For example, when the screen size was increased from 10,000 to 20,000 proteins per round, no further improvement in the final evolved energies was seen. Interestingly, the optimal simulated DNA shuffling length of 20 amino acids (60 bases) is nearly identical to fragment lengths used in experimental protocols (Crameri *et al.,* 1998).

3. Single-Pool Swapping

In Nature, local protein space can be rapidly searched by the directed recombination of encoded domains from multigene pools. A prominent example is the creation of the primary antibody repertoire in an adaptive immune system. These events are generalized by simulating the swapping of amino acid fragments from five structural pools representing helices, strands, loops, turns, and others (see Fig. 6c). During the swapping step, subdomains were randomly replaced with members of the same secondary structural pools with an optimal probability of 0.01/subdomain/round. The simulated evolution of the primary fold is limited by maintaining the linear order of swapped secondary structure types. The addition of the swapping move was so powerful that it was possible to achieve binding constants 2 orders of magnitude higher than in shuffling simulations (see Table III). Significantly, these improved binding constants were achieved starting with 10–20 times less minimized structural subdomain material.

4. Mixing

Parallel tempering is a powerful statistical method that often allows a system to escape local energy minima (Geyer, 1991). This method simultaneously simulates several systems at different temperatures, allowing systems at adjacent temperatures to swap configurations. The swapping between high- and low-temperature systems allows for an effective searching of configuration space. This method achieves rigorously correct canonical sampling, and it significantly reduces the equilibration time in a simulation. Instead of a single system, a larger ensemble with n systems is considered in parallel tempering, and each system is equilibrated at a distinct temperature T_i, $i = 1, \ldots, n$. The procedure in parallel tempering is illustrated in Fig. 7.

The system with the lowest temperature is the one of our interest; the higher-temperature systems are added to aid in the equilibration of the system of interest. In addition to the normal Monte Carlo moves performed in each system, swapping moves are proposed that exchange the configurations between two systems i and $j = i + 1$, $1 \leq i < n$. The higher-temperature systems are included solely to help the lowest-temperature system to escape from local energy minima via the swapping moves. To achieve efficient sampling, the highest temperature should be such that no significant free energy barriers are observed. So that the swapping moves are accepted with a reasonable probability, the energy histograms of systems adjacent in the temperature ladder should overlap.

In Nature, as well, it is known that genes, gene fragments, and gene operons are transfered between species of different evolutionary complexity (i.e., at different "temperatures"). By analogy, limited population mixing is performed among several parallel swapping experiments by randomly

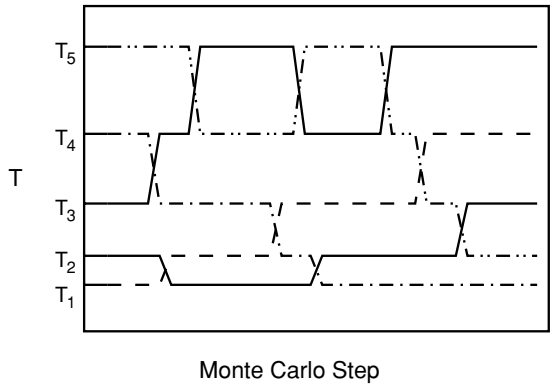

FIG. 7. A schematic drawing of the swapping taking place during a parallel tempering simulation (from Falcioni and Deem, 1999).

exchanging evolving proteins at an optimal probability of 0.001/protein/round. These mixing simulations optimized local space searching and achieved binding constants $\approx 10^5$ higher than did base substitution alone (see Table III). Improved function is due, in part, to the increased number of events in parallel experiments. Indeed, mixing may occur in Nature when the evolutionary target function changes with time. That is, in a dynamic environment with multiple selective pressures, mixing would be especially effective when the rate of evolution of an isolated population is slower than the rate of environmental change. It has also been argued that spatial heterogeneity in drug concentration, a form of "spatial parallel tempering," facilitates the evolution of drug resistance (Kepler and Perelson, 1998).

5. Multipool Swapping

The effective navigation of protein space requires the discovery and selection of tertiary structures. To model the large-scale search of this space, a random polypeptide sequences was used as a starting point, and the swapping protocol was repeated. Now, however, secondary structures from all five pools were permitted to swap in at every subdomain (see Fig. 6d). This multipool swapping approach evolved proteins with binding constants $\approx 10^7$ better than did amino acid substitution of a protein-like starting sequence (see Table III). This evolution was accomplished by the random yet correlated juxtaposition of different types of low-energy secondary structures. This approach dramatically improved specific ligand binding while efficiently discovering new tertiary structures (see Fig. 8). Optimization of the rate of these hierachical molecular evolutionary moves, including relaxation of the selection criteria, enabled the protein to evolve despite the high rate of failure for these dramatic swapping moves. Interestingly, of all the molecular evolutionary processes modeled, only multipool swapping demonstrated chaotic behavior in repetitive simulations. This chaotic behavior was likely due to the discovery of different model folds that varied in their inherent ability to serve as scaffolds for ligand specific binding.

F. Possible Experimental Implementations

An important motivation of this work was that the proposed protocols must be experimentally feasible. Indeed, the ultimate test of the effectiveness of these protocols will be experimental. It is hoped that the search of large regions of protein space apparently possible with these methods will identify new protein folds and functions of great value to basic, industrial, and medical research.

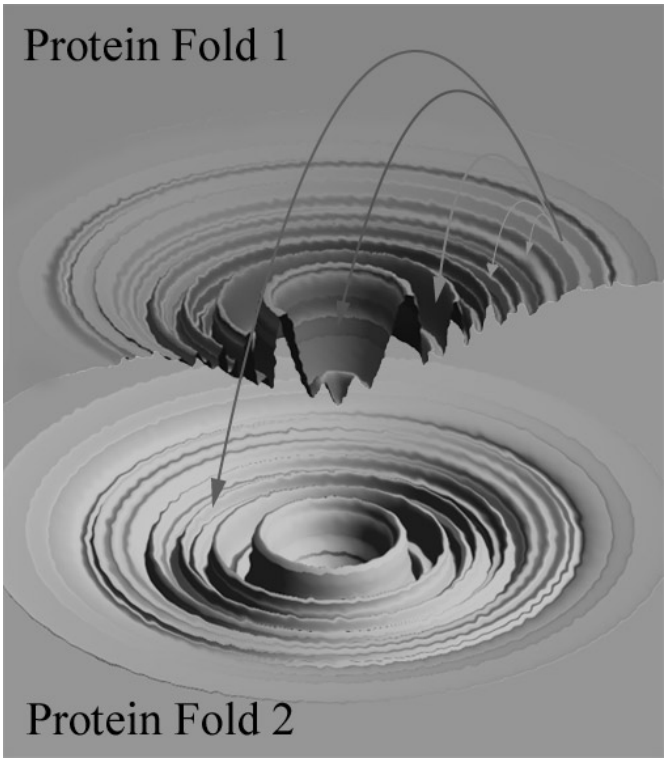

FIG. 8. Schematic diagram representing a portion of the high-dimensional protein composition space (from Bogarad and Deem, 1999). The three-dimensional energy landscape of protein fold 1 is shown in cutaway. The arcs with arrowheads represent the ability of a given molecular evolution process to change the composition and so to traverse the increasingly large barriers in the energy function. The smallest arc represents the ability to evolve improved fold function via point mutation; then, in increasing order, DNA shuffling, swapping, and mixing. Finally, the multipool swapping protocol allows an evolving system to move to a different energy landscape representing a new tertiary fold (bottom).

The main technical challenge posed by the swapping protocols is the nonhomologous recombination necessary to swap the DNA that codes for different secondary structures into the evolving proteins. One approach would be to generate multiple libraries of synthetic oligonucleotide pools (Mandecki, 1990; Stemmer *et al.,* 1995) encoding the different secondary subdomain structures. Asymmetric, complementary encoded linkers with embedded restriction sites would make the assembly, shuffling, and swapping steps possible.

Alternatively, the techniques of ITCHY (Ostermeier *et al.,* 1999a) and SCRATCHY (Ostermeier *et al.,* 1999b) may be used to accomplish the nonhomologous swapping of secondary structures required within our protocol.

Finally, exon shuffling may be used to perform the nonhomologous swapping events. In this case, the pools of secondary structures would be encoded with exons of a living organism, such as *E. coli.* There is precedent for such use of exon shuffling, at both the DNA (Fisch *et al.,* 1996) and the RNA (Moran *et al.,* 1999) level.

G. Life Has Evolved to Evolve

Although the focus has been on the higher levels of the evolution hierarchy, because that is where the biggest theoretical and experimental gap lies, all levels are important. In particular, the details of how point mutation assists protein evolution are important.

DNA base mutation leads only indirectly to changes in protein expression. How mutations occur in the bases and how these mutations lead to codon changes, and so amino acid changes, is not purely random. The inherent properties of the genetic code and biases in the mechanisms of DNA base substitution are perfectly suited for the "neutral" search of local space. Previously, the genetic code has been presented as a nodal or hypercubic structure to illustrate these relationships (Maeshiro and Kimura, 1998; Jiménez-Montaño *et al.,* 1996). It seems preferable to view the standard genetic code quantitatively as a 64×64 two-dimensional matrix. Seeds of this approach can be found in Kepler's work regarding evolvability in immunoglobulin genes (Cowell *et al.,* 1999; Kepler and Bartl, 1998; Kepler, 1997). The values in this matrix are the probabilities of a specific codon mutating to another by a single base change under error-prone conditions, e.g., mutator strains of bacteria, error-prone PCR, or somatic hypermutation. Assuming each base mutates independently in the codon, this matrix can be calculated from a simpler 4×4 matrix of base mutation probabilities. The base mutation matrix can be extracted from available experimental data (Smith *et al.,* 1996). A synonymous transition probability can be defined for each codon, which is the probability per replication of a base change that leads to a codon that codes for the same amino acid. A conservative transition probability can further be defined, which is the probability per replication that a base change leads to a conservative mutation. Finally, a nonconservative transition probability can be defined, which is the probability per replication that a base change leads to a nonconservative mutation. The conservative and nonconservative mutation probabilities can be viewed as defining the evolutionary potential of each codon: codons with high conservative and nonconservative mutation rates can be said to exhibit a high evolutionary potential. These mutation tendencies are shown in Fig. 9. In general, amino acids that exhibit a dramatic functional property, such as the charged residues, the ringed residues, cysteine, and tryptophan, tend to mutate at higher nonconservative rates

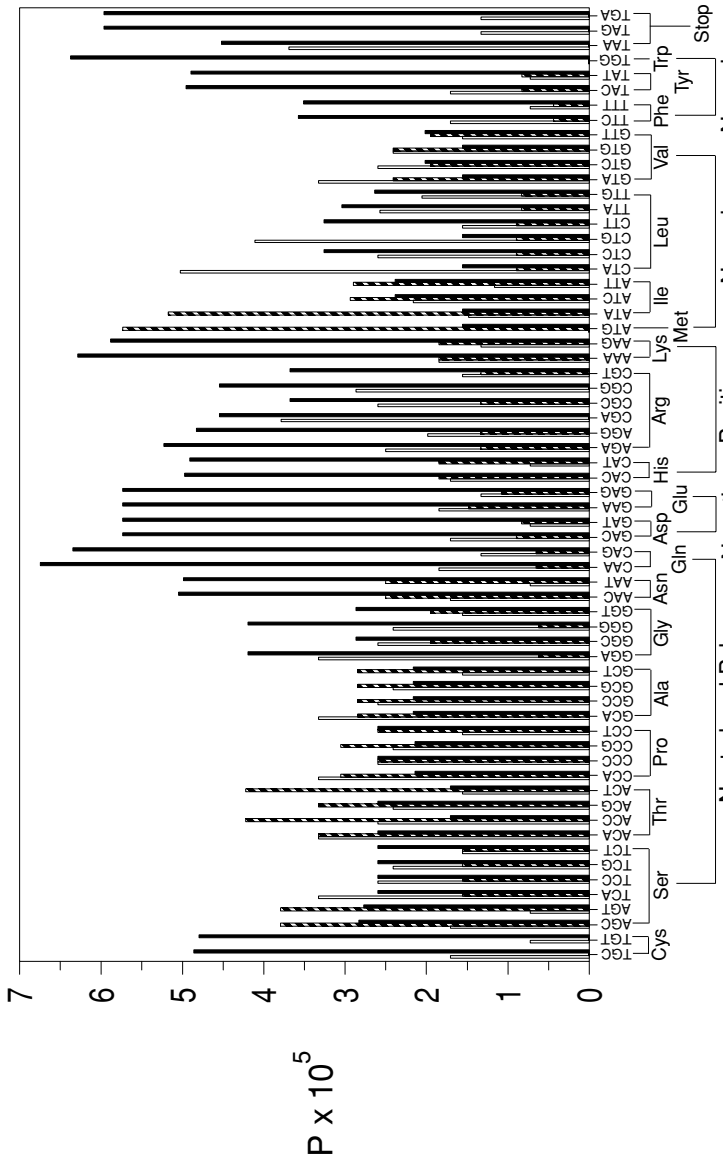

FIG. 9. Shown are the probabilities of a given codon mutating in a synonymous (open), conservative (hatched), and non-conservative (filled) way in one round of replication under error-prone conditions. The codons are grouped by the amino acid encoded, and the amino acids are grouped by category.

that allow for the possible deletion of the property. Amino acids that are more generic, such as the polar neutral residues and the nonpolar nonringed residues, tend to mutate at higher conservative rates that allow sequence space to be searched for similar favorable contacts.

There is a connection between condon usage and DNA shuffling. In the most successful DNA shuffling experiments by the Stemmer group, codon assignments of the initial coding sequences are optimized for expression. This assignment typically increases the nonsynonymous mutation rates described above. In particular, this assignment tends to increase the conservative mutation rate for the generic amino acids and the nonconservative mutation rate for the dramatically functional amino acids. Codon usage, then, has already implicitly been used to manipulate mutation rates. Explicit consideration of the importance of codon assignment would be an interesting amino acid level refinement of existing molecular evolution protocols. Optimized protocol parameters can be identified, taking into account the detailed codon usage information. Similarly, the codon potential matrix can be used in the design of the pools of secondary structures in the swapping-type molecular evolution protocols. That is, DNA can be chosen that codes for the secondary structures that (i) tends not to mutate, (ii) tends to mutate to synonymous sequences, (iii) tends to mutate to conservative sequences, or (iv) tends to mutate to non-conservative sequences.

In Nature there are numerous examples of exploiting codon potentials in ongoing evolutionary processes (Kepler, 1997). In the V regions of encoded antibodies, high-potential serine codons such as AGC are found predominantly in the encoded CDR loops, while the encoded frameworks contain low-potential serine codons such as TCT. Unfortunately, antibodies and drugs are often no match for the hydrophilic, high-potential codons of "error-prone" pathogens. The dramatic mutability of the HIV gp120 coat protein is one such example. One can envision a scheme for using codon potentials to target disease epitopes that mutate rarely (i.e., low-potential) and unproductively (i.e., become stop, low-potential, or structure-breaking codons). Such a therapeutic scheme is quite simple, and so could be quite generally useful against diseases that otherwise tend to become drug resistant.

H. Natural Analogs of These Protocols

During the course of any evolutionary process, proteins become trapped in local energy minima. Dramatic moves, such as swaps and juxtaposition, are needed to break out of these regions. Dramatic moves are usually deleterious, however. The evolutionary success of these events depends on population size, generation time, mutation rate, population mixing, selective

pressure or freedom, such as successful genome duplications or the establishment of set-aside cells (Davidson *et al.,* 1995), and the mechanisms that transfer low-energy, encoded structural domains.

By using the analogy with Monte Carlo to design "biased" moves for molecular evolution, a swapping-type move has been derived that is similar to several mechanisms of natural evolution. Viruses and transposons, for example, have evolved large-scale integration mechanisms (Pennisi, 1998). Exon shuffling is also a generator of diversity, and a possible scenario is that exon shuffling generated the primordial fold diversity (Gilbert, 1978; Gilbert *et al.,* 1997; Netzer and Hartl, 1997). Alternatively, random swapping by horizontal transfer (Lawrence, 1997), rearrangement, recombination, deletion, and insertion can lead to high in-frame success rates in genomes with high densities of coding domains and reading frames, as in certain prokaryotes and mitochondria.

While inter- and intraspecies exchange of DNA is often thought to occur primarily on the scale of genes and operons, shorter exchange often occurs. Indeed, the most prevalent exchange length within *E. coli* is of the order of several hundred to a thousand base pairs (Syvanen, 1997). Similarly, analysis of the evolution of vertebrate cytochrome *c* suggests that transfer of segments significantly smaller than a single gene must have occurred (Syvanen, 1997).

Indeed, swapping mutations leading to significant diversity are not rare in Nature. *Neisseria meningitidis* is a frequent cause of meningitis in sub-Saharan Africa (Hobbs *et al.,* 1997). The Opa proteins are a family of proteins that make up part of the outer coat of this bacteria. These proteins undergo some of the same class switching and hypervariable mutations as antibody domains. A significant source of diversity also appears to have come from interspecies transfer with *Neisseria gonorrhoeae* (Hobbs *et al.,* 1997). Both the surface coat proteins and the pilon proteins of *N. gonorrhoeae* undergo significant homologous recombination to produce additional diversity. It appears generally true that the intra- and interspecies transfer of short segments of genes is common in *E. coli, Streptococci,* and *Neisseria*. Rapid evolution of diversity such as this obviously poses a significant challenge for therapeutic protocols.

Three dramatic examples of use of swapping by Nature are particularly notable. The first is the development of antibiotic resistance. It was originally thought that no bacteria would become resistant to penicillin due to the many point mutations required for resistance. Resistance occurred, however, within several years. It is now known that this resistance occurred through the swapping of pieces of DNA between evolving bacteria (Shapiro, 1992, 1997). One mechanism of antibiotic resistance was incorporation of genes coding for β-lactamases. These genes, which directly degrade β-lactam antibiotics,

appear to be relatively ancient, and their incorporation is a relatively simple example of a swapping-type event. These β-lactamases have continued to evolve, however, in the presence of antibiotic pressure, both by point mutation and by shuffling of protein domains via exon shuffling (Maiden, 1998; Medeiros, 1997). The bacterial targets of penicillin, the penicillin-binding proteins, are modular proteins that have undergone significant structural evolution since the introduction of penicillin (Massova and Mobashery, 1999; Goffin and Ghuysen, 1998). This evolution was of a domain-shuffling form, and it is a more sophisticated example of a Natural swapping-type move. Multidrug resistance is, of course, now a major, current health care problem. The creation of the primary antibody repertoire in vertebrates is another example of DNA swapping (of genes, gene segments, or pseudo-genes). Indeed, the entire immune system mostly likely evolved from a single transposon insertion some 450 million years ago (Plasterk, 1998; Agrawal *et al.,* 1998). This insertion, combined with duplication and subsequent mutation of a single membrane spanning protein, mostly likely lead to the class switching apparatus of the primary repertoire. Finally, the evolution of *E. coli* from *Salmonella* occurred exclusively by DNA swapping (Lawrence, 1997). None of the phenotypic differences between these two species is due to point mutation. Moreover, even the observed rate of evolution due to DNA swapping, 31,000 bases/million years, is higher than that due to point mutation, 22,000 bases/million years. Even though a DNA swapping event is less likely to be tolerated than is a point mutation, the more dramatic nature of the swapping event leads to a higher overall rate of evolution. This is exactly the behavior observed in the simulated molecular evolutions.

I. Concluding Remarks on Molecular Evolution

DNA base substitution, in the context of the genetic code, is ideally suited for the generation, diversification, and optimization of local protein space (Miller and Orgel, 1974; Maeshiro and Kimura, 1998). However, the difficulty of making the transition from one productive tertiary fold to another limits evolution via base substitution and homologous recombination alone. Nonhomologous DNA recombination, rearrangement, and insertion allow for the combinatorial creation of productive tertiary folds via the novel combination of suitable structures. Indeed, efficient search of the high-dimensional fold space requires a hierarchical range of mutation events.

This section has addressed from a theoretical point of view the question of how protein space can be searched efficiently and thoroughly, either in the laboratory or in Nature. It was shown that point mutation alone is incapable of evolving systems with substantially new protein folds. It was further

demonstrated that even the DNA shuffling approach is incapable of evolving substantially new protein folds. The Monte Carlo simulations demonstrated that nonhomologous DNA "swapping" of low-energy structures is a key step in searching protein space.

More generally, the simulations demonstrated that the efficient search of large regions of protein space requires a hierarchy of genetic events, each encoding higher order structural substitutions. It was shown how the complex protein function landscape can be navigated with these moves. It was concluded that analogous moves have driven the evolution of protein diversity found in Nature. The proposed moves, which appear to be experimentally feasible, would make an interesting addition to the techniques of molecular biology. An especially important application of the theoretical approach to molecular evolution is modeling the molecular evolution of disease.

There are many experimental applications of the technology for molecular evolution (Patten *et al.,* 1997). Perhaps some of the most significant are in the field of human therapeutics. Molecular evolution can be used directly to improve the performance of protein pharmaceuticals. Molecular evolution can be used indirectly to evolve small molecule pharmaceuticals by evolving the pathways that code for small molecule synthesis in *E. coli.* Molecular evolution can be used for gene therapy and DNA vaccines. Molecular evolution can be used to produce recombinant protein vaccines or viral vaccines. Finally, molecular evolution can be used to create modified enzymatic assays in drug screening efforts. The ability to develop new assays that do not infringe on competitors' techniques is an important ability for large pharmaceutical companies, given the current complex state of patent claims. There is a similar range of applications of molecular evolution in the field of biotechnology. As shown in Table II, many of the tools of molecular biology can be improved or modified through the use of molecular evolution.

A wide variety of pest organisms and parasites, including fungi, weeds, insects, protozoans, macroparasites, and bacteria, have used evolutionary processes to evade chemical control. The range of evolutionary events exhibited by these organisms is similar in spirit to the hierarchy of moves present in the molecular evolution protocol (see Fig. 6). Bacteria provide one of the most pressing examples of the problems posed by an evolving disease (a "moving target"). Although there undoubtedly have been many selective pressures upon bacteria, the novel pressure with the largest impact in the last half-century has been the worldwide use of antibiotics. This background presence of antibiotics has led to the development of antibiotic resistance in many species of bacteria. Indeed, multidrug resistance is now a major health care issue, with some strains resistant to all but one, or even all, known antibiotics.

Interestingly, there is another strong pressure on evolving bacteria, that of the vertebrate immune system. This pressure is thought to be responsible for mosaic, or modular as a result of swapping-type events, genes found in species of bacteria not naturally genetically competent, such as *E. coli* and *S. pyogenes* (Dowson *et al.,* 1997). In these cases, the long-standing, strong selective pressure due to the interaction with the immune system likely led to genetic exchange.

Kepler and Perelson (1998) have noted that a spatial heterogeneity in the concentration of a drug can facilitate evolved resistance in a disease organism. This occurs because regions of low drug concentration provide a "safe harbor" for the disease, where replication and mutation can occur. The regions of high disease concentration provide the selective pressure for the evolution. Explicit examples of this mode of evolution include the role of spatial heterogeneity in the spread of insecticide resistance, noncompliance to antibiotic regimes in the rise of resistance in the tuberculosis bacterium, and heterogeneity within the body of the protease inhibitor indinavir in the rise of resistant HIV-1 strains (Kepler and Perelson, 1998). As noted above, this type of evolution is a spatial example of parallel tempering, a technique that has proven to be very powerful at sampling difficult molecular systems with many and large energy barriers. This analogy with parallel tempering suggests that heterogeneities must be of great and intrinsic importance in natural evolution.

Qualitative changes in protein space such as those modeled here allow viruses, parasites, bacteria, and cancers to evade the immune system, vaccines, antibiotics, and therapeutics. All of these pathogens evolve, to a greater or lesser degree, by large, swapping-type mutations. The successful design of vaccines and drugs must anticipate the evolutionary potential of both local and large space searching by pathogens in response to therapeutic and immune selection. The addition of disease specific constraints to simulations such as these should be a promising approach for predicting pathogen plasticity. Indeed, infectious agents will continue to evolve unless we can force them down the road to extinction.

IV. Summary

Significant opportunities exist for the application of ideas from statistical mechanics to the burgeoning area of combinatorial chemistry. While combinatorial chemistry was not invented by researchers in the field of statistical mechanics, it is fair to say that perhaps it should have been! The design of effective experimental methods for searching composition space is similar in

concept to the design of effective Monte Carlo methods for searching configuration space. Optimization of the parameters in combinatorial chemistry protocols is analogous to the integration of various types of moves in Monte Carlo simulation. It is notable that one of the strongest present chemical-engineering proponents of combinatorial chemistry in the solid state, Henry Weinberg at SYMYX Technologies, has taught graduate statistical mechanics for the last 25 years! Applications of combinatorial chemistry abound in the fields of catalysis, sensors, coatings, microelectronics, biotechnology, and human therapeutics. Hopefully, statistical mechanics will have a significant role to play in shaping these new methods of materials design.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge the contributions of my collaborators Leonard D. Bogarad, Marco Falcioni, and Taison Tan. This work was supported by the National Science Foundation.

REFERENCES

Agrawal, A., Eastman, Q. M., and Schatz, D. G., Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **394,** 744–751 (1998).

Akporiaye, D. E., Dahl, I. M., Karlsson, A., and Wendelbo, R., Combinatorial approach to the hydrothermal synthesis of zeolites. *Angew. Chem. Int. Ed.* **37,** 609–611 (1998).

Bogarad, L. D., and Deem, M. W., A hierarchical approach to protein molecular evolution. *Proc. Natl. Acad. Sci. USA* **96,** 2591–2595 (1999).

Bratley, P., Fox, B. L., and Niederreiter, H., Algorithm-738-programs to generate Niederreiter's low-discrepancy sequences. *ACM Trans. Math. Software* **20,** 494–495 (1994).

Briceño, G., Chang, H., Sun, X., Schultz, P. G., and Xiang, X.-D., A class of cobalt oxide magnetoresistance materials discovered with combinatorial synthesis. *Science* **270,** 273–275 (1995).

Burgess, K., Lim, H.-J., Porte, A. M., and Sulikowski, G. A., New catalysts and conditions for a C-H insertion reaction identified by high throughput catalyst screening. *Angew. Chem. Int. Ed.* **35,** 220–222 (1996).

Cole, B. M., Shimizu, K. D., Krueger, C. A., Harrity, J. P. A., Snapper, M. L., and Hoveyda, A. H., Discovery of chiral catalysts through ligand diversity: Ti-catalyzed enantioselective addition of TMSCN to *meso* epoxides. *Angew. Chem. Int. Ed.* **35,** 1668–1671 (1996).

Cong, P., Doolen, R. D., Fan, Q., Giaquinta, D. M., Guan, S., McFarland, E. W., Poojary, D. M., Self, K., Turber, H. W., and Weinberg, W. H., High-throughput synthesis and screening of combinatorial heterogeneous catalyst libraries. *Angew. Chem. Int. Ed.* **38,** 484–488 (1999).

Cowell, L. G., Kim, H. J., Humaljoki, T., Berek, C., and Kepler, T. B., Enhanced evolvability in immunoglobulin V genes under somatic hypermutation. *J. Mol. Evol.* **49,** 23–26 (1999).

Crameri, A., Raillard, S. A., Bermudez, E., and Stemmer, W. P. C., DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391,** 288–291 (1998).

Danielson, E., Devenney, M., Giaquinta, D. M., Golden, J. H., Haushalter, R. C., McFarland, E. W., Poojary, D. M., Reaves, C. M., Weinberg, W. H., and Wu, X. D., A rare-earth phosphor containing one-dimensional chains identified through combinatorial methods. *Science* **279,** 837–839 (1988).

Danielson, E., Golden, J. H., McFarland, E. W., Reaves, C. M., Weinberg, W. H., and Wu, X. D., A combinatorial approach to the discovery and optimization of luminescent materials. *Nature* **389,** 944–948 (1997).

Davidson, E. H., Peterson, K. J., and Cameron, R. A., Origin of bilaterian body plans—Evolution of developmental regulatory mechanisms. *Science* **270,** 1319–1325 (1995).

de Pablo, J. J., Laso, M., and Suter, U. W., Estimation of the chemical potential of chain molecules by simulation. *J. Chem. Phys.* **96,** 6157 (1992).

Dickinson, T. A., and Walt, D. R., Generating sensor diversity through combinatorial polymer synthesis. *Anal. Chem.* **69,** 3413–3418 (1997).

Dowson, C. G., Barcus, V., King, S., Pickerill, P., Whatmore, A., and Yeo, M., Horizontal gene transfer and the evolution of resistance and virulence determinants in *Streptococcus*. *J. Appl. Micro. Biol. Symps. Supl.* **83,** 42S–51S (1997).

Falcioni, M., and Deem, M. W., A biased Monte Carlo scheme for zeolite structure solution. *J. Chem. Phys.* **110,** 1754–1766 (1999).

Falcioni, M., and Deem, M. W., Library design in combinatorial chemistry by Monte Carlo methods. *Phys. Rev. E* **61,** 5948–5952 (2000).

Fisch, I., Kontermann, R. E., Finnern, R., Hartley, O., Solergonzalez, A. S., Griffiths, A. D., and Winter, G., A strategy of exon shuffling for making large peptide repertoires displayed on filamentous bacteriophage. *Proc. Natl. Acad. Sci. USA* **93,** 7761–7766 (1996).

Frenkel, D., and Smit, B., Unexpected length dependence of the solubility of chain molecules. *Mol. Phys.* **75,** 983 (1992).

Frenkel, D., and Smit, B., "*Understanding Molecular Simulation: From Algorithms to Applications*." Academic Press, San Diego, 1996.

Frenkel, D., Mooij, C. G. A. M., and Smit, B., Novel scheme to study structural and thermal properties of continuously deformable molecules. *J. Phys. Condens. Matter* **4,** 3053 (1992).

Geyer, C. J., Markov chain Monte Carlo maximum likelihood. In "Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface." American Statistical Association, New York, 1991, pp. 156–163.

Gilbert, W., Why genes in pieces? *Nature* **271,** 501 (1978).

Gilbert, W., DeSouza, S. J., and Long, M., Origin of genes. *Proc. Natl. Acad. Sci. USA* **94,** 7698–7703 (1997).

Goffin, C., and Ghuysen, J.-M., Multimodular penicillin-binding proteins: An enigmatic family of orthologs and paralogs. *Microbiol. Mol. Biol.* **62,** 1079–1093 (1998).

Helmkamp, M. M., and Davis, M. E., Synthesis of porous silicates. *Annu. Rev. Mater. Sci.* **25,** 161–192 (1995).

Hobbs, M. M., Seiler, A., Achtman, M., and Cannon, J. G., Microevolution within a clonal population of pathogenic bacteria: Recombination, gene duplication and horizontal genetic exchange in the *opa* gene family of *Neisseria meningitidis*. *Mol. Microbiol.* **12,** 171–180 (1997).

Jiménez-Montaño, M. A., de la Mora-Basánez, C. R., and Pöschel, T., The hypercube structure of the genetic code explains conservative and non-conservative amino acid substitutions *in vivo* and *in vitro*. *BioSystems* **39,** 117–125 (1996).

Kamtekar, S., Schiffer, J. M., Xiong, H. Y., Babik, J. M., and Hecht, M. H., Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262,** 1680–1685 (1993).

Kauffman, S. A., "The Origins of Order." Oxford University Press, New York, 1993.

Kauffman, S. A., and MacReady, W. G., Search strategies for applied molecular evolution. *J. Theor. Biol.* **173,** 427–440 (1995).

Kauffman, S., and Levin, S., Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128,** 11–45 (1987).

Kepler, T. B., Codon bias and plasticity in immunoglobulins. *Mol. Biol. Evol.* **14,** 637–643 (1997).

Kepler, T. B., and Bartl, S., Plasticity under somatic mutation in antigen receptors. *Curr. Topics Micrbiol.* **229,** 149–162 (1998).

Kepler, T. B., and Perelson, A. S., Drug concentration heterogeneity facilitates the evolution of drug resistance. *Proc. Natl. Acad. Sci. USA* **95,** 11514–11519 (1998).

Lawrence, J. G., Selfish operons and speciation by gene transfer. *Trends Microbiol.* **5,** 355–359 (1997).

Maeshiro, T., and Kimura, M., Role of robustness and changeability on the origin and evolution of genetic codes. *Proc. Natl. Acad. Sci. USA* **95,** 5088–5093 (1998).

Maiden, M. C. J., Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacertia. *Clin. Infect. Dis.* **27,** S12–S20 (1998).

Mandecki, W., A method for construction of long randomized open reading frames and polypeptides. *Protein Eng.* **3,** 221–226 (1990).

Marinari, E., Parisi, G., and Ruiz-Lorenzo, J., Numerical simulations of spin glass systems. In *Spin Glasses and Random Fields,* "Directions in Condensed Matter Physics" (A. Young, ed.), World Scientific, Singapore, Vol. 12, 1998, pp. 59–98

Massova, I., and Mobashery, S., Structural and mechanistic aspects of evolution of β-lactamases and penicillin-binding proteins. *Curr. Pharm. Design* **5,** 929–937 (1999).

Medeiros, A. A., Evolution and dissemination of β-lactamases accelerated by generation of β-lactam antiobiotics. *Clin. Infect. Dis.* **24,** S19–S45 (1997).

Menger, F. M., Eliseev, A. V., and Migulin, V. A., Phosphatase catalysis develped *via* combinatorial organic chemistry. *J. Org. Chem.* **60,** 6666–6667 (1995).

Miller, S., and Orgel, L., "The Origin of Life on Earth." Prentice Hall, London, 1974.

Moore, J. C., Jin, H.-M., Kuchner, O., and Arnold, F. H., Strategies for the *in-vitro* evolution of protein function—Enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.* **272,** 336–347 (1997).

Moran, J. V., DeBerardinis, R. J., and Kazazian, H. H., Exon shuffling by L1 retrotransposition. *Science* **283,** 1530 –1534 (1999).

Netzer, W. J., and Hartl, F. U., Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* **388,** 343–349 (1997).

Niederreiter, H., "Random Number Generation and Quasi-Monte Carlo Methods," Society for Industrial and Applied Mathematics, Philadelphia, 1992.

Novet, T., Johnson, D. C., and Fister, L., Interfaces, interfacial reactions and superlattice reactants. *Adv. Chem. Ser.* **245,** 425–469 (1995).

Ostermeier, M., Nixon, A. E., and Benkovic, S. J., Incremental truncation as a strategy in the engineering of novel biocatalysts. *Bioorg. Med. Chem.* **7,** 2139–2144 (1999a).

Ostermeier, M., Shim, J. H., and Benkovic, S. J., A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature Biotech.* **17,** 1205–1209 (1999b).

Patten, P. A., Howard, R. J., and Stemmer, W. P. C., Applications of DNA shuffling to pharmaceuticals and vaccines. *Curr. Opin. Biotech.* **8,** 724–733 (1997).

Pennisi, E., How the genome readies itself for evolution. *Science* **281,** 1131–1134 (1998).

Perelson, A. S., and Macken, C. A., Protein evolution on partially correlated landscapes. *Proc. Natl. Acad. Sci. USA* **92,** 9657–9661 (1995).

Pirrung, M. C., Spatially addressable combinatorial libraries. *Chem. Rev.* **97,** 473–488 (1997).

Plasterk, R., V(D)J recombination: Ragtime jumping. *Nature* **394,** 718–719 (1998).

Reddington, E., Sapienza, A., Gurau, B., Viswanathan, R., Sarangapani, S., Smotkin, E. S., and Mallouk, T. E., Combinatorial electrochemistry: A highly parallel, optical screening method for discovery of better electrocatalysts. *Science* **280,** 1735–1737 (1998).

Riddle, D. S., Santiago, J. V., Brayhall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q., and Baker, D., Functional rapidly folding proteins from simplified amino acid sequences. *Nature Struct. Biol.* **4,** 805–809 (1997).

Schuster, P., and Stadler, P. F., Sequence redundancy in biopolymers: A study on RNA and protein structures. In *Viral Regulatory Structure and Their Degeneracy* (G. Myers, ed.), Addison–Wesley, New York, 1998, pp. 163–186.

Sedgewick, R., "Algorithms" 2nd ed., Addison–Wesley, New York, 1998.

Shapiro, J. A., Natural genetic engineering in evolution. *Genetica* **86,** 99–111 (1992).

Shapiro, J. A., Genome organization, natural genetic engineering and adaptive mutation. *Trends Genet.* **13,** 98–104 (1997).

Smith, B., and Maesen, T. L. M., Commensurate 'freezing' of alkanes in the channels of a zeolite. *Nature* **374,** 42 (1995).

Smith, D. S., Creadon, G., Jena, P. K., Portanova, J. P., Kotzin, B. L., and Wysocki, L. J., Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. *J. Immunol.* **156,** 2642–2652 (1996).

Stemmer, W. P. C., Rapid evolution of a protein *in-vitro* by DNA shuffling. *Nature* **370,** 389–391 (1994).

Stemmer, W. P. C., Crameri, A., Ha, K. D., Brennan, T. M., and Heyneker, H. L., Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164,** 49–53 (1995).

Syvanen, M., Horizontal gene transfer: Evidence and possible consequences. *Annu. Rev. Genet.* **28,** 237–261 (1997).

van Dover, R. B., Schneemeyer, L. F., and Fleming, R. M., Discovery of a useful thin-film dielectric using a composition-spread approach. *Nature* **392,** 162–164 (1998).

Volkenstein, M. V., "Physical Approaches to Biological Evolution." Springer-Verlag, New York, 1994.

Wang, J., Yoo, Y., Gao, C., Takeuchi, I., Sun, X., Chang, H., Xiang, X.-D., and Schultz, P. G., Identification of a blue photoluminescent composite material from a combinatorial library. *Science* **279,** 1712–1714 (1998).

Weinberg, W. H., Jandeleit, B., Self, K., and Turner, H., Combinatorial methods in homogeneous and heterogeneous catalysis. *Curr. Opin. Chem. Biol.* **3,** 104–110 (1998).

Xiang, X.-D., Sun, X., Briceño, G., Lou, Y., Wang, K.-A., Chang, H., Wallace-Freedman, W. G., Chang, S.-W., and Schultz, P. G., A combinatorial approach to materials discovery. *Science* **268,** 1738–1740 (1995).

Zhang, J.-H., Dawes, G., and Stemmer, W. P. C., Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. *Proc. Natl. Acad. Sci. USA* **94,** 4504–4509 (1997).

Zones, S. I., Nakagawa, Y., Lee, G. S., Chen, C. Y., and Yuen, L. T., Searching for new high silica zeolites through a synergy of organic templates and novel inorganic conditions. *Micropor. Mesopor. Mat.* **21,** 199–211 (1998).