Information Directed Sampling for Combinatorial Material Synthesis and Library Design

CHIA HUANG YEN, DAVID SHAN HILL WONG
AND SHI SHANG JANG
Department of Chemical Engineering,
National Tsing Hua University, Hsinchu, Taiwan 30043

***Keywords*:** Information Theory, Combinatorial Methods, Generalized Regression, Neural Network

Combinatorial techniques have become more and more important in many areas of chemistry and chemical engineering research. It was suggested that simulated annealing can be used to improve the efficiency of sampling in combinatorial methods. However, without priori model estimates of fitness function, true importance sampling cannot be performed. In this case, the efficiency of annealing is only as good as random search. We suggested that a simple prediction model using currently available data can be constructed using a generalized regression neural network. An index of our uncertainty about a point in the search space can also be established using information entropy. An information free energy combined the two indices to direct the search so that importance sampling is performed. Two benchmark problems were used to model the optimization problem involved in combinatorial synthesis and library design. We showed that when importance sampling is performed, the combinatorial technique became much more effective. The improvement in efficiency over undirected methods is especially significant when the size of the problem becomes very large.

Introduction

In recent years, combinatorial synthesis become an important optimization technique in product and process development (Davis, 1999), e.g. material synthesis (Xiang *et al.*, 1995; Szostak, 1997; Wilson and Czarnik, 1997; Danielson *et al.*, 1998; Klein *et al.*, 1998; van Dover *et al.*, 1998; Cong *et al.* 1999; Engstrom and Weinberg, 2000), design of catalysts (Cole *et al.*, 1996; Jandeleit *et al.*, 1998; Schlögl, 1998; Senkan, 1998), selection of solvent (Pretel *et al.*, 1994), drug design (Gordon *et al.*, 1996; Gordon, 1998; Linusson *et al.*, 2000), improvement of enzymes activity (You and Arnold 1996; Bornscheuer, 1998). Combinatorial techniques screen a large quantity of different combinations of inputs to find the condition that produces the best merit function.

There are two focuses in combinatorial synthesis research: (1) how to create a large throughput of experiments (e.g. Hanak, 1970), (2) how to develop a protocol for designing experiments, so that the input state-space can be sampled effectively (e.g. Gordon, 1998; Linusson *et al.*, 2000; Voigt *et al.*, 2001). In practice, the two problems may not be separable since the method of experiments may dictate under what conditions a batch of experiments may be done. The selection of experiments is then highly correlated instead of random. However, in this work, we are concerned only with how previous information obtained can be used to improve sampling efficiency.

Sampling policy in combinatorial synthesis is just another form of experimental design. In the past, we have proposed experimental design methods based on information theory (Lin *et al.*, 1995; Chen *et al.*, 1998) for process optimization and recipe selection. The philosophy is as follows. Given a set of data, an empirical model can be trained. This model can be used to direct the search and suggest experiments at points that have potentially high fitness function (low information energy). However, such a model is untrustworthy when data are sparse compared to the entire search space (high information entropy). We need to explore areas that have not yet been investigated. A temperature-annealing schedule can be used to shift from an information-entropy-based search to information-energy-based search. These methods are not efficient for combinatorial synthesis because the modeling techniques used are not equipped to handle problems of extremely large dimensional. The neural network used require extensive training. In this work, we shall propose an information-based importance sampling policy that can be used for combinatorial synthesis. The efficiency of this method is tested using the RPV and the N–K models.

Received on October 25, 2002. Correspondence concerning this article should be addressed to S.-S. Jang (E-mail address: ssjang@che.nthu.edu.tw).

1. Theory

1.1 Fitness landscape

1.1.1 N–K model There are two types of optimization problems in combinatorial chemistry: high-dimensional "structural" search of combinations of integer variables and high dimensional "spatial" search of continuous variables.

In structural search problems, components in a library are combined to form a particular structure. Experiments or calculations are then carried out to see if the synthesized structure has the desired property. A benchmark problem of such optimization is known as the N–K problem (Kauffman and Levin, 1987). The N–K model captures the basic physics of many phenomena such as genomics, protein evolution, etc. (Kauffman, 1993; Perelson and Macken, 1995). Various modified forms of the N–K model have also been proposed (Bogarad and Deem, 1999). The simplest N–K model can be described as follows. Consider an N-dimensional array of integer variables $(a_1, a_2, \cdots, a_N)$. It represents the state of a sequence. Each entry of the sequence a_j represents a component from the library. For example, in gene sequencing, each entry a_j can be one of the four nucleotides C, T, G, A. In protein folding problems, each entry a_j can be a particular form of amino acids. In drug design, each entry a_j can be one of the functional groups that are used to form a ligand. In the N–K problem, a_j is a binary variable that takes values of 1 or 0. The merit function $E(a_1, a_2, \ldots, a_N)$ is given by:

$$E(a_1, a_2, \cdots, a_N) = \frac{1}{[(N-K)]^{1/2}} \sum_{j=1}^{N=N-K+1} \sigma_j(a_j, a_{j+1}, \cdots, a_{j+K-1}) \qquad (1)$$

$\sigma_j(a_j, a_{j+1}, \cdots, a_{j+K-1})$ is the contribution to the merit function by the j-th entry. Note that the contribution depends not only on the state of occupancy of the j-th variable, but also on how the next $K-1$ variables are occupied. Such a formulation is an analog of many physical situations. For example, a particular phenotype is found only if a particular sequence of nucleotides (a genotype) is found. In a protein folding problem, the state of a particular amino-acid in a protein depends on its neighboring amino acids. A functional group manifests its chemical nature only if the neighboring groups provide a suitable molecular environment. In the N–K problem $\sigma_j(a_j, a_{j+1}, \cdots, a_{j+k-1})$ is given by a "lookup" table of problem-defined values. Two examples of lookup tables for ($N = 4$, $K = 2$) and ($N = 4$, $K = 3$) are given in **Fig. 1**. Given the lookup tables, the merit function of all 16 different states can be calculated. For example, in Fig. 1, given the lookup table ($N = 4$, $K = 2$), the point (0, 0, 1, 0) will have a

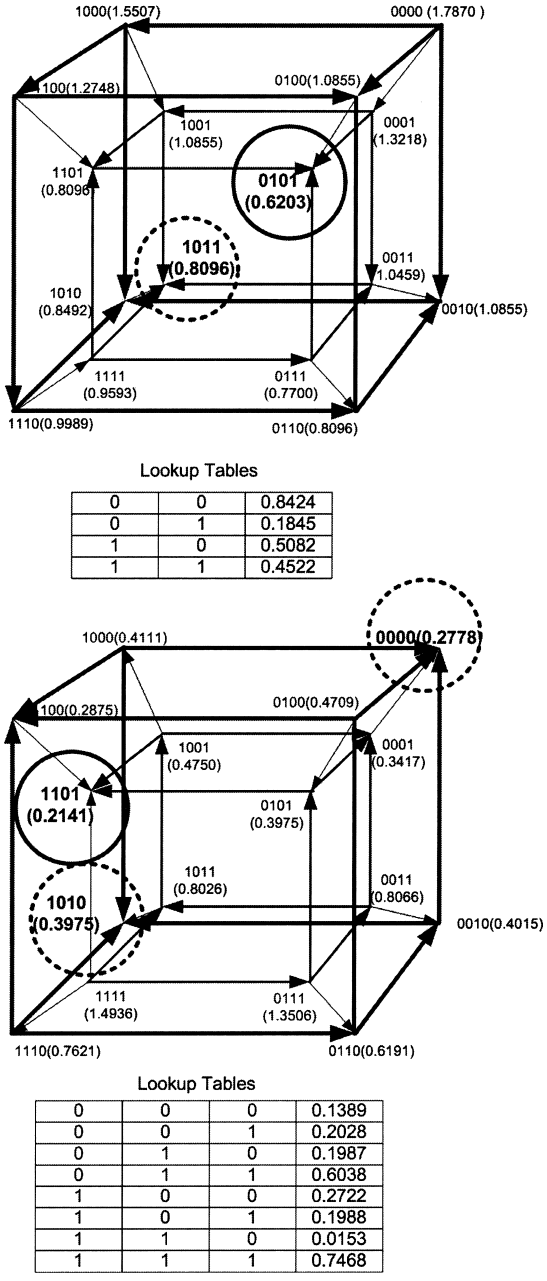


Fig. 1 Local and global extremas for NK models

merit function

$$E_{1000} = \left(\sqrt{4-2}\right)^{-1}\left(\sigma(1,0)+\sigma(0,0)+\sigma(0,0)\right)$$
$$= \left(\sqrt{2}\right)^{-1}(0.5082+0.8424)$$
$$= 1.5507$$

For $N = 4$, the changes of the merit function can be shown as two connecting Boolean cubes. Each vertex is connected to 4 different vertices, each of which differs from the center point by one variable. For example, the point (1, 0, 0, 0) is connected to (0, 0, 0, 0),

(1, 1, 0, 0), (1, 0, 1, 0) and (1, 0, 0, 1). The connections can be shown as edges of the connecting cubes. The arrows show the direction of decrease in the merit function. The lookup table (N = 4, K = 2) generates two local minima, while the other with N = 4, K = 3 generates three. Kauffman (1993) showed that as N and K increase, the number of local minima increases, and the difference between global optimal and average value of the merit function decreases. For more details of the N–K model and its relevance to various physical situations, readers can refer to the book by Kauffman (Kauffman, 1993). The role of the N–K problem for library design optimization is analogous to that of the traveling salesman problem for network routing.

1.1.2 Random phase volume model In many combinatorial synthesis problems, experimental variables include the compositions of the material and processing conditions. The input state-space consists of both discrete and continuous variables and is typically of very high dimensions. The merit function is usually one or a set of specific properties of the material such as superconductivity, luminescence, catalytic activity, tensile strength, etc. Such properties will depend on the particular phase of the product material. Since the change of physical property of the material is discontinuous across a phase boundary, the objective function encountered in combinatorial synthesis optimization is only piecewise continuous.

Falcioni and Deem (2000) proposed a "random phase volume" (RPV) model to simulate the merit function encountered in combinatorial synthesis. Essentially RPV is a relation between the merit function E and a set of compositional variables $\vec{x}$ and non-composition variables $\vec{z}$. Composition variables are expressed as mole fractions. Therefore for a C component system, there are $C-1$ independent composition variables. The entire composition space is divided into $\alpha = 1, \cdots, M$ different phases by defining M phase center points $\vec{x}_\alpha^0$. Each point in the composition space belongs to the phase of the nearest phase center point. Non-composition variables may be processing conditions such as temperature, pressure, pH, time of reaction, thickness of film, etc. They may be discrete or continuous and normalized in the range of [–1, 1] in the RPV model. Similarly the non-composition space is divided into $\gamma = 1, \cdots, N$ different phases by defining N phase center points $\vec{z}_\alpha^0$:

$$E(\vec{x},\vec{z}) = E_{\alpha\gamma}$$
$$\text{if} \quad \left\|\vec{x}-\vec{x}_\alpha^0\right\| = \min_{\alpha'=1,\cdots,M}\left\|\vec{x}-\vec{x}_{\alpha'}^0\right\|$$
$$\text{and} \quad \left\|\vec{z}-\vec{z}_\gamma^0\right\| = \min_{\gamma'=1,\cdots,M}\left\|\vec{z}-\vec{z}_{\gamma'}^0\right\| \tag{2a}$$

The merit function $E_{\alpha\gamma}$ in this phase is represented by a

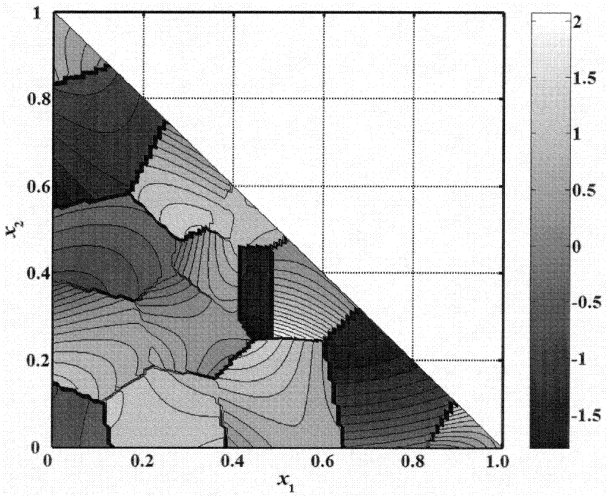


Fig. 2 A typical merit function landscape of the random-phase volume model

Q_x-th order polynomial in composition variable and a Q_z-th order polynomial in non-composition variable:

$$E_{\alpha\gamma} = U_\alpha + \sigma_x \sum_{k=1}^{Q_x} \sum_{i_1 \geq \cdots \geq i_k = 1}^{C} f_{i_1 \cdots i_k} \times \xi_x^{-k} A_{i_1 \cdots i_k}^{(\alpha k)} y_{i_1} y_{i_2} \cdots y_{i_k} + \frac{1}{2}\left(W_\gamma + \sigma_z \sum_{k=1}^{Q_z} \sum_{i_1 \geq \cdots \geq i_k = 1}^{D} f_{i_1 \cdots i_k} \times \xi_z^{-k} B_{i_1 \cdots i_k}^{(\gamma k)} w_{i_1} w_{i_2} \cdots w_{i_k} \right) \tag{2b}$$

with $\vec{y} = \vec{x} - \vec{x}_\alpha$ and $\vec{w} = \vec{z} - \vec{z}_\alpha$. U_α, W_γ, $A_{i_1 \cdots i_k}^{(\alpha k)}$, $B_{i_1 \cdots i_k}^{(\gamma k)}$, are parameters generated by Gaussian random number with zero mean and unit variance. The scale factors ξ_x and ξ_z are chosen so that each term in the multinomial expansion contributes roughly the same amount. The σ_x and σ_z are chosen so that the multinomial expansion terms contribute about 40% of $E_{\alpha\gamma}$. The symmetric factor is given by:

$$f_{i_1 \cdots i_k} = \frac{k!}{\prod_{i=1}^{l} o_i!} \tag{2c}$$

where l is the number of distinct integer values in the set of $\{i_1, \cdots, i_k\}$, 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 $\sum_{i=1}^{l} o_{i=k}$. **Figure 2** is a typical merit function landscape of an RPV model with C = 3, D = 0, M = 15 and Q_x = 2. The piecewise continuous and nonlinear nature is evident. The complexity of the problem can be increased by increasing the dimension of variable and number of phases. One can appreciate the similarity of RPV model landscape in Fig. 2 by comparing

with the merit function landscape obtained in actual combinatorial synthesis study (e.g. Cong *et al*., 1999).

Given the N–K model and the RPV model, the sampling algorithm for optimization in combinatorial synthesis and library design can be benchmarked.

1.2 Sampling policy

The simplest way of sampling the state space is to use grid search followed by local search. However, grid search becomes inefficient as the dimensionality of state-space increases. Another class of search method commonly used to solve high dimensional optimization is stochastic search methods such as simulated annealing (SA, Kirkpatrick *et al.*, 1983). In a typical SA scheme applied to combinatorial synthesis, a set of reference data points are created; a set of candidate points are generated. The values of the merit function of the candidate data points are compared to those of the reference data points. The reference data point is then updated by the candidate according to the transition probability:

$$P = \begin{cases} 1 & Y_i^{\mathrm{c}} < Y_i^{\mathrm{r}} \\ \exp\left(\dfrac{-\left(Y_i^{\mathrm{c}} - Y_i^{\mathrm{r}}\right)}{T}\right) & Y_i^{\mathrm{c}} \geq Y_i^{\mathrm{r}} \end{cases} \tag{3}$$

where Y_i^{c} is the merit function of candidate data and Y_i^{r} is the merit function of reference data. The temperature is reduced gradually according to a predetermined annealing schedule to ensure that the system is not trapped in a local minimum.

The updating procedure, given in Eq. (3), mimics the importance-sampling strategy, known as the Metropolis scheme, used in Monte Carlo simulation of molecular ensembles. Importance sampling is pivotal in calculating ensemble average in Monte Carlo simulation because it reduces the number of times that total energy of unimportant states are calculated and sampled in the ensemble average. The number of interactions evaluated in calculating the total energy of the ensemble is proportional to $n(n - 1)/2$, n being the number of molecules. The number of interactions evaluated in calculating the transition probability in Eq. (3) is proportional to n. Thus importance sampling reduces the computing time significantly but ensures that states sampled in the ensemble average are distributed according to the Maxwell–Boltzmann distribution.

The objective of the combinatorial synthesis is, however, slightly different. It is desirable that the true global minimum is located, not the true ensemble average. Moreover, the merit function of a data point is known only if the actual experiment is performed. Therefore, unless some prior estimates of the merit function are available, importance sampling cannot be performed. The act of updating according to the difference between the merit function of the candidate and the reference point according to the Metropolis scheme only prevents the search from being trapped in a local minimum prematurely. Without real importance sampling, simulated annealing will not be any more effective than random search.

1.3 Generalized regression neural network

In order that existing data can be efficiently modeled, the generalized regression network was used (Specht, 1991). If the joint probability density function $p(Y, \vec{X})$ of a random input variable at $\vec{X}$ having an output variable Y is known, then the expected value of the output at $\vec{X}$ is given by:

$$\left\langle Y\left(\vec{X}\right)\right\rangle = \frac{\int_{-\infty}^{\infty} Y p\left(Y, \vec{X}\right) dY}{\int_{-\infty}^{\infty} p\left(Y, \vec{X}\right) dY} \tag{4}$$

If $p(Y, \vec{X})$ is unknown, it can be estimated using existing measurements ($\vec{X}^i$, Y^i), $i = 1, \cdots, n$, as Gaussian distribution function (Parzen, 1962):

$$\tilde{p}\left(Y, \vec{X}\right) = \frac{1}{(2\pi)^{(p+1)/2}\sigma^{(p+1)}}\frac{1}{n} \cdot \sum_{i=1}^{n} \exp\left[-\frac{\left(\vec{X} - \vec{X}^i\right)^T\left(\vec{X} - \vec{X}^i\right)}{2\sigma^2}\right] \exp\left[\frac{\left(Y - Y^i\right)^2}{2\sigma^2}\right] \tag{5}$$

Expected value of the output at $\vec{x}$ is given by

$$\left\langle \tilde{Y}\left(\vec{X}\right)\right\rangle = \frac{\sum_{i=1}^{n} \exp\left[-\dfrac{\left(\vec{X} - \vec{X}^i\right)^T\left(\vec{X} - \vec{X}^i\right)}{2\sigma^2}\right] Y^i}{\sum_{i=1}^{n} \exp\left[-\dfrac{\left(\vec{X} - \vec{X}^i\right)^T\left(\vec{X} - \vec{X}^i\right)}{2\sigma^2}\right]} \tag{6}$$

There are several advantages of using GRNN in our method.

1. GRNN requires no training. The prediction model of GRNN in Eq. (6) requires only ($\vec{X}^i$, Y^i), the location and results of data already sampled. σ is a problem-specific smoothing factor which is preset and kept constant. Therefore, the training phase of GRNN is only a one-pass reading of all existing data.

2. Both integer and continuous inputs can be used in GRNN although different values of σ may be

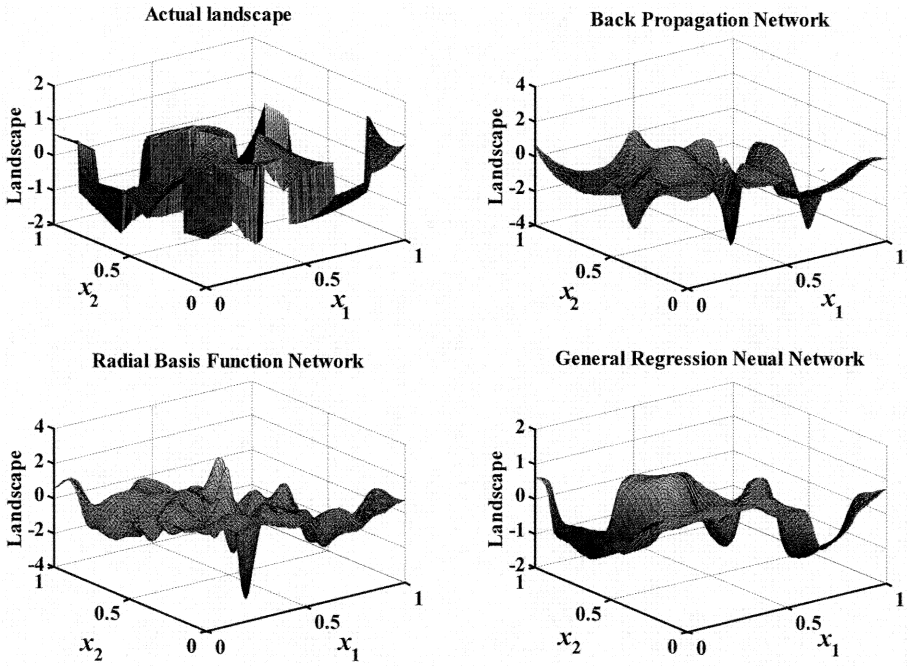


Fig. 3 Fitness function surface obtained using different neural networks for RPV models

needed. Therefore N–K landscape can be modelled efficiently. We compared the training and prediction accuracies of a GRNN model to another popular neural network, the radial basis function network (RBFN, Chen *et al.*, 1991) for a small N(=10)–K(=5) model. The model has 1024 states. 100 states were randomly taken as the training data, and the rest were chosen as test set data. While both models can accurately reproduce the training data, the prediction mean square error of GRNN was found to be 0.066, much less than that found by RBFN (0.519). Obviously, GRNN produces a much better model when the amount of data sample is limited.

3. The predictions of GRNN always fall within the maximum and minimum of the training data set. Use of GRNN avoids unreasonable extrapolation. This is extremely important if we wish to model a piecewise continuous function such as the RPV model. **Figure 3** compared GRNN, RBFN and back-propagation network (BPN; Haykin, 1999) modeling of an RPV model with a limited amount of training data. Again, GRNN produces the most reasonable fitness landscape for the RPV model.

1.4 Information energy

Given a model of all the presently available experimental data, we define the information energy index of any candidate point in the search space

$$U(\vec{x}) = \begin{cases} \left\langle \tilde{Y}(\vec{X}) \right\rangle & \text{for minimization problems} \\ -\left\langle \tilde{Y}(\vec{X}) \right\rangle & \text{for maximization problems} \end{cases} \tag{7}$$

where $<\tilde{Y}(\vec{X})>$ indicates the predicted value of the fitness function using a GRNN model. At any candidate point, the better the predicted fitness function, the lower is its information energy. The more valuable is the information at that point.

1.5 Information entropy

Our knowledge of a candidate point can be measured by the information entropy (Shannon, 1948),

$$S(\vec{X}) = \frac{\int_{-\infty}^{\infty} p(Y,\vec{X}) \ln p(Y,\vec{X}) dY}{\int_{-\infty}^{\infty} p(Y,\vec{X}) dY} \tag{8}$$

Given the distribution function in Eq. (5), the information entropy can be estimated as:

$$S(\vec{X}) = \frac{\sum_{i=1}^{n} \frac{(\vec{X}-\vec{X}^i)^T(\vec{X}-\vec{X}^i)}{2\sigma^2} \exp\left[-\frac{(\vec{X}-\vec{X}^i)^T(\vec{X}-\vec{X}^i)}{2\sigma^2}\right]}{\sum_{i=1}^{n} \exp\left[-\frac{(\vec{X}-\vec{X}^i)^T(\vec{X}-\vec{X}^i)}{2\sigma^2}\right]} \tag{9}$$

The information entropy is just the average of the square of the distance between a candidate point and all existing data points. The higher the information entropy, the more we know about the candidate point.

1.6 Information free energy and temperature annealing

A candidate is worthy of experiment if it has a potential of having a good fitness value (low information energy), or its neighborhood has not been sufficiently explored (high information entropy). During the initial stages when the number of experiment is small, the model predictions are of little value, experiments should be devoted to sample un-chartered search space. When sufficient information has been gathered, only samples that are potentially important should be tested. Chen *et al.* (1998) proposed an information free energy index:

$$F = U - TS \tag{10}$$

with T being an annealing temperature proportional to the number of experiment. The proper convergence of the sampling procedure to the global minimum depends on the annealing schedule. In this work, the very fast re-annealing scheme proposed by Ingber (1989) was used:

$$T = T_0 \cdot \exp(-c \cdot k^{1/m}) \tag{11}$$

with T_0, c, m being adjustable constants and k is the total number of the experiments.

1.7 Flowchart

Our proposed sampling policy can be summarized into a flowchart in **Fig. 4** as the following steps.

(i) Obtain a GRNN model for all existing data.

(ii) Select the data with the best fitness function as a reference.

(iii) Use random search to produce a candidate experiment. Check if the experiment has been performed. Generate another candidate if this experiment has been performed. If the experiment has not been performed, proceed to step (iv).

(iv) Use GRNN to calculate information energy and information free energy at the candidate point.

(v) If the information free energy is smaller than or equal to the free energy of the existing reference data, check if the batch has been filled. If not go to step (iii), to generate additional experiments.

(vi) If the information free energy is greater than the free energy of the existing reference data, check if the number of trial exceeds a preset number NT (NT should be very large). If the number of trial is less than NT, go to step (iii). If the number has reached NT, a local search is used to create a new experiment.

(vii) Repeat steps (iii) to (vi) until all candidate experiments of the new batch have been selected.

(viii) Perform the experiments.

(ix) Repeat step (i) to (viii) at a reduced temperature until some stopping criterion is met, i.e. the fitness function has met some specifications or the number of batches reaches a preset number NB.

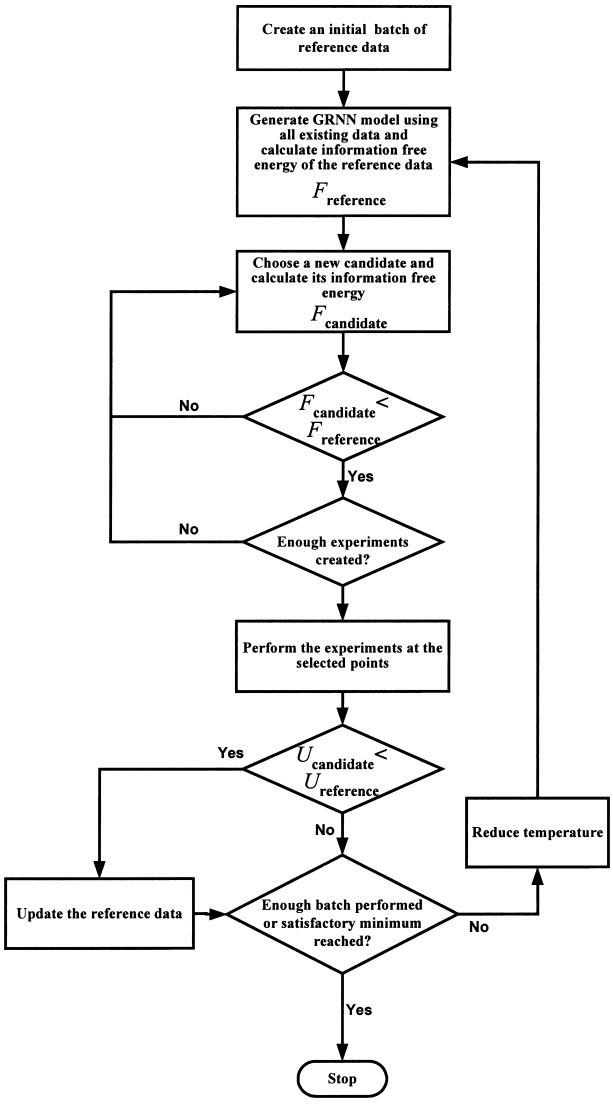


Fig. 4 Flow chart of information directed search

We shall call the above sampling policy "free energy directed simulated annealing" (Free Energy-DSA, FEDSA).

In the above algorithm, a local search is used if directed search cannot produce an acceptable candidate. In the N–K model, this is done by mutating a small number of variables of the array. In RPV model, this is done by taking a random small step away from the reference minimum.

Two undirected search methods: random search (RS) and undirected simulated annealing (USA) are used for comparison. In RS, a set of candidate experi-

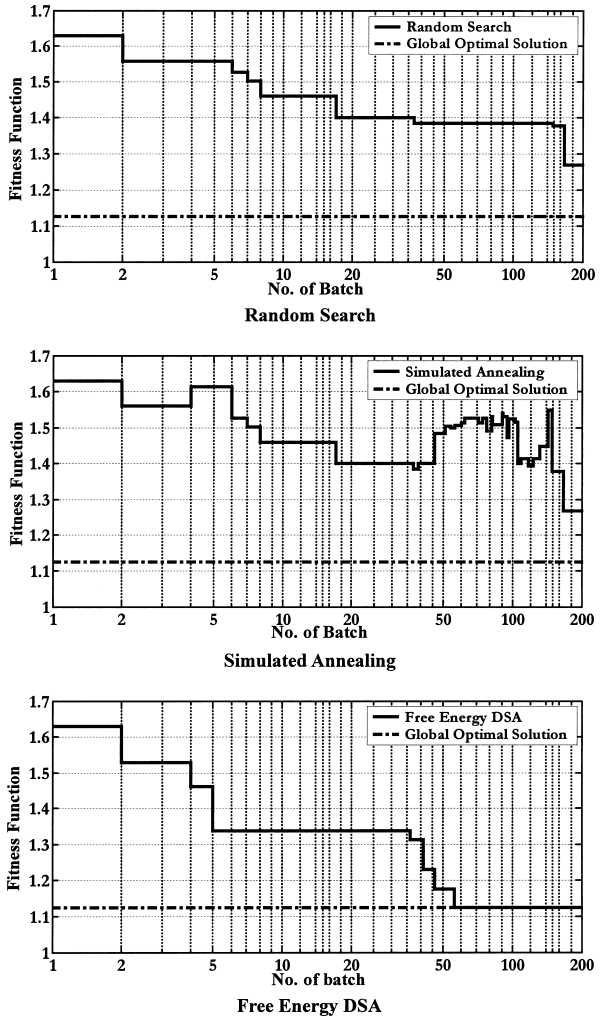


Fig. 5 Changes in the best fitness function located using different search strategies for a simple N–K problem ($N = 20$, $K = 5$)

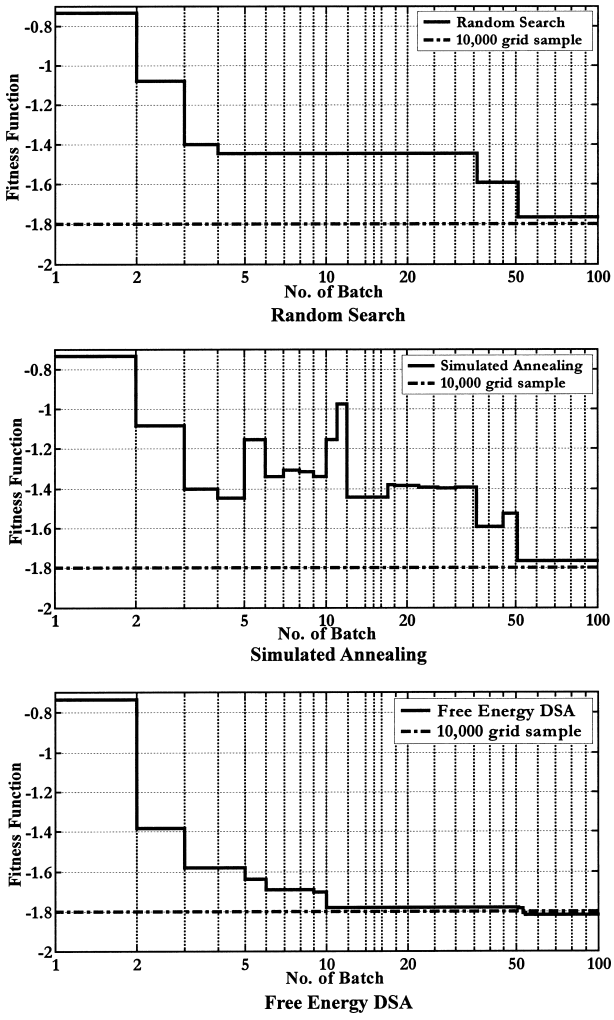


Fig. 6 Changes in the best fitness function located using different search strategies for a simple RPV problem ($C = 3$, $M = 15$, $Q_x = 2$)

ments are selected randomly. The experiments are performed and the samples in the reference data set are updated if the fitness value of the new experiments is better than that of the reference data. In undirected simulated annealing (USA), a random search generates the new experiment candidates. The experiments are performed without further screening. The Metropolis scheme in Eq. (3) is used to determine whether the reference data is replaced by the new data using experimental results.

2. Results and Discussion

2.1 Search efficiency

To demonstrate the advantages of our approach, we shall use two simple examples: (i) an N–K model with $N = 20$, Boolean state inputs of 0 and 1 and $K = 4$, (ii) the RPV model with $C = 3$, $D = 0$, $M = 15$ and $Q_x = 2$. The small N–K problem has only 1,048,576 states that allow exhaustive search and identification of the true global minimum and the true fitness function distribution. The two-dimensional nature of the RPV problem allows visualization of the search process.

Figures 5 and **6** illustrate the decrease in the objective function with the number of batches of experiments for the small N–K model and the RPV model respectively. For a pure random search, the fitness function initially decreases rapidly but improvement becomes more and more difficult. An undirected simulated annealing search causes the system to deviate from the current minimum so that the search process will avoid being trapped in a local minimum. Without importance sampling, generation of states with better objective fitness is by pure chance. The rate of finding the best fitness function can only be as fast as random search. If a model of the fitness function is constructed using previous data and importance sampling (directed

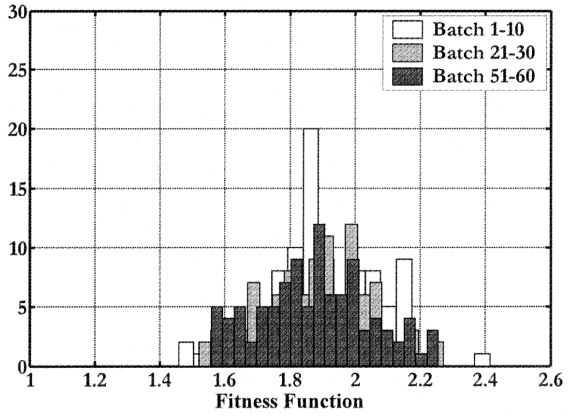
annealing) are performed based on information free energy, the rate of finding the best fitness function can be much improved. In Fig. 5, we found that the FEDSA located true global minimum even though the landscape of the N–K model is very rugged. Random search and undirected simulated annealing fails to locate the true minimum. Similarly, FEDSA is able to find the approximate region of the true optimum at around the 10th batch for the RPV model (Fig. 6). At this point, the values of the fitness function obtained by a random search procedure and simulated annealing are much higher. Random search and undirected simulated annealing fail to locate the true minimum.

2.2 Importance sampling

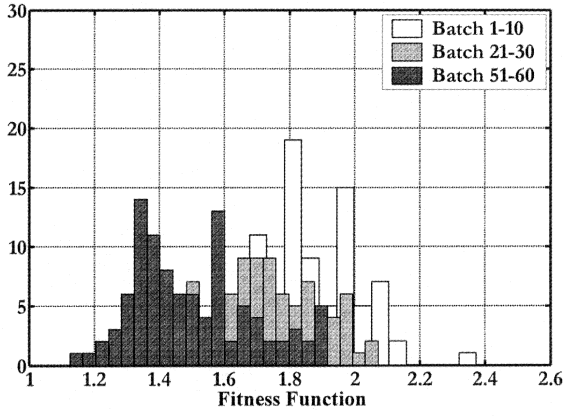
Figures 7 and **8** illustrate the distribution of the data sampled at different stages in optimization. For random search the distribution is similar to simulated annealing and therefore not shown in the figures.

Figure 7 shows the distribution of the data sampled from the 1st to 10th batches, 21st to 30th batches and 51st to 60th batches. It is obvious that the distribution of the data sampled shifted towards low energy end. Such changes were not found for simulated annealing. FEDSA allows us to perform importance sampling in the early stages of the search. After the 50th batch, mutation becomes dominant. There is not much shift in the energy distribution of the data sampled. During the later stages of the search, mutation frequently became the sampling technique rather than importance sampling. While the data sampled differ in only a few bits, the rugged Nature of the fitness landscape results in a broad distribution of fitness values for the sampled data.

For the RPV model, local search is important even when temperature annealing virtually stopped. Importance sampling allows us to locate the correct phase. Figure 8 illustrates the distributions of the data during different stages of the optimization. In FEDSA, data are scattered across the state space initially. Then data are selected around the projected local minima. Finally all data are concentrated in the correct phase with the global minimum. This again illustrates the efficiency of importance sampling. On the other hand, the data are scattered around the entire search space during different stages of the optimization for undirected simulated annealing. No importance sampling was performed. The backgrounds of the figures in the column under undirected simulated annealing in Fig. 8 are the true contours of the RPV model. The backgrounds of the figures under free energy DSA are contours of the GRNN model. Note that the GRNN model captured the general feature of the RPV model, but the details are far from the global optimum. When we attempt to create a model of a large search space it is important that efforts are not wasted in reproducing the exact details in regions that are not relevant in an optimization sense.



(a) Undirected Simulated Annealing



(b) Free Energy DSA

Fig. 7 Distribution of data sampled by different search strategies at different stages of optimization for the N–K model

2.3 Effect of problem size

If the size of the N–K model increases, exhaustive search becomes difficult. Similarly as the size of the RPV model increases visualization becomes impossible. **Table 1** presents the optimal fitness function located for a given number of experiments (NB = 100). Since these are stochastic searches, the optimal value located will vary from run to run as the size of the problem increases. Therefore the averages and standard deviations of 10 different runs are listed. It was found that Free-Energy-DSA always has lower averages and smaller standard deviations than simulated annealing. The difference between FEDSA and USA/RS becomes larger as the size of the problem increases. This emphasizes the need of true importance sampling in solution of complex problems.

2.4 Computation Effort

The main objective of our approach is to reduce the number of actual sampling. In this benchmarking study, an actual sampling is equivalent to a function evaluation. Figures 5 and 6 compared the values of objective function obtained at different number of func-

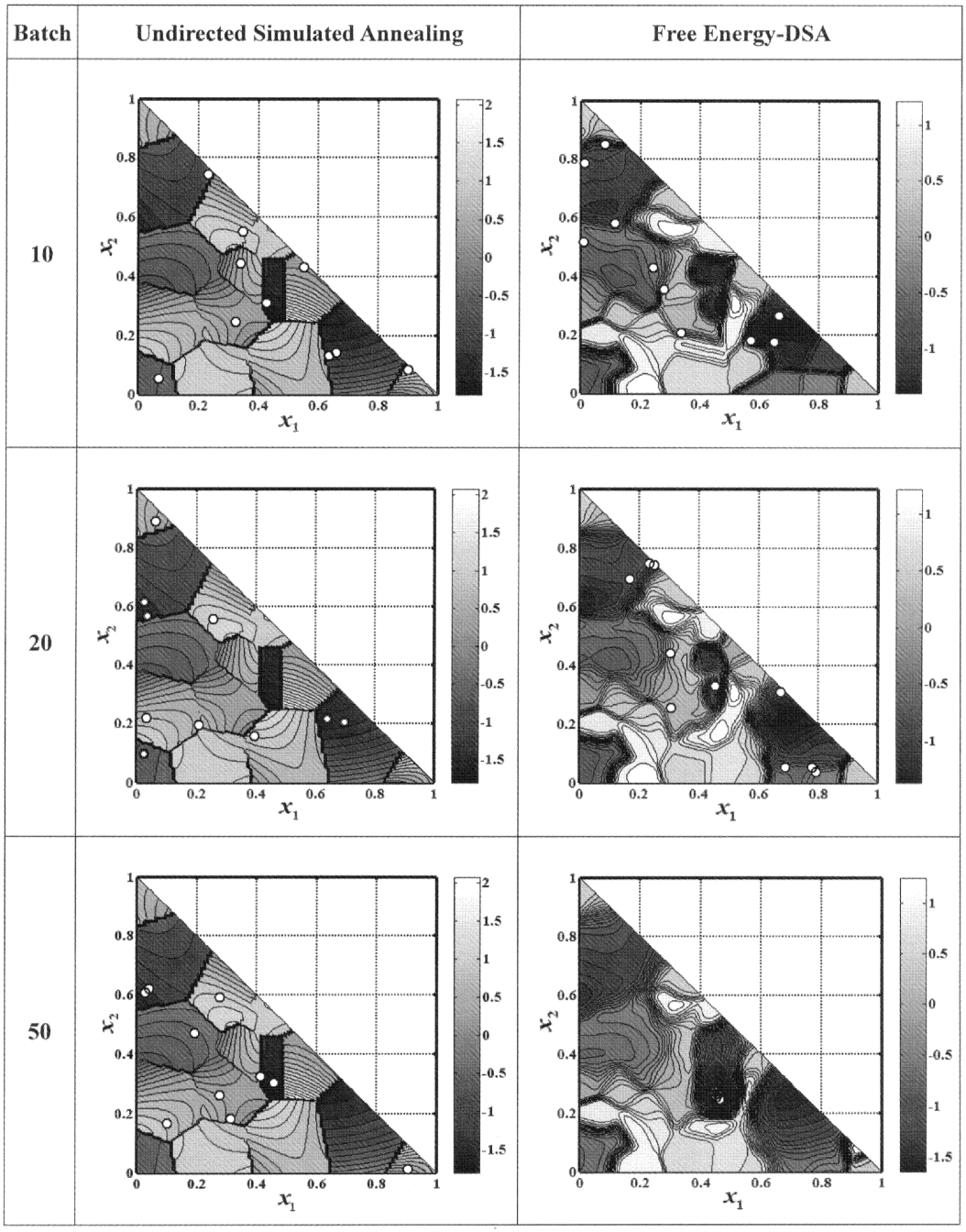


Fig. 8 Distribution of sampling points at different stages of optimization for the RPV model(C = 3, M = 15, Q_x = 2)

tion evaluations. It is obvious that our information directed approach is more efficient than an undirected sampling method such as random search or simulated annealing in terms of finding an optimum landscape with a least number of samplings. Since GRNN requires no training, only a little extra computation effort is required to determine which sample is worth testing. In the simulation study using N–K and RPV, since evaluation of the objective function is relatively easy, the net saving of computation time is negligible. However, an actual library design and combinatorial synthesis, an actual sampling usually involves the performance of an experiment, or running an elaborate molecular simulation. Actual sampling are time consuming and expensive. The advantage of information-directed sampling will be substantial.

Conclusions

In this work, we have demonstrated the importance of true importance sampling in solving optimization problems of high dimension. Our results show that

Table 1 Optimal obtained for a fixed number of experiment (100 batch of 10 experiment each)

	Undirected simulated annealing		Free-energy DSA	
	Mean	Standard deviation	Mean	Standard deviation
RPV model				
$M=15$, $C=3$, $D=0$, $Q_x=2$	–1.71	0.08	–1.827	0.006
$M=15$, $C=3$, $D=0$, $Q_x=6$	–1.24	0.16	–1.697	0.004
$M=37$, $C=6$, $D=0$, $Q_x=2$	–1.01	0.21	–1.987	0.052
$M=37$, $C=6$, $D=0$, $Q_x=6$	–1.02	0.12	–1.618	0.022
NK model				
$N=20$, $K=4$	1.54	0.08	1.13	0
$N=25$, $K=5$	1.65	0.14	1.39	0.03
$N=30$, $K=5$	1.88	0.11	1.58	0.07

brute force combinatorial techniques may be powerful, but the technique becomes much more effective if we can organize the present knowledge periodically to direct the search. We have demonstrated that the organization of knowledge need not be done in a theoretical manner. It can be done by constructing a simple empirical model with sufficient flexibility. Due to the large amount of data involved, this model should require a minimal regression computation but must also be statistically sound. A generalized regression neural network was selected to perform the task of modeling. Furthermore, during early stages of the search, we must not put too much emphasis on model predictions since a large fraction of the search space remains unexplored. An information entropy index allows us to direct the search to unexplored regions of the search space. An information free energy index was used to balance the need to confirm the model predictions of regions of optimality and the need of chartering unexplored search space. True importance sampling can be achieved using this information free energy index and effectiveness of combinatorial can be substantially improved.

Acknowledgment

The authors thank the financial support provided by the National Science Council, Taiwan, for this work through the grant NSC90-2622-E007-003.

Nomenclature

$A_{i_1\cdots i_k}^{(\alpha k)}$ = parameters in the PRV model
a_j = j-th entry of the N-dimensional array in the NK model
$B_{i_1\cdots i_k}^{(\gamma k)}$ = parameters in the PRV model
C = dimensions of the composition variables in the RPV model
c = parameters in very fast simulated re-annealing
D = dimensions of the non-composition variables in the RPV model
E = merit function of the RPV model
$E_{\alpha\gamma}$ = merit function of the NK model
F = information free energy
f_i = symmetry factor of the polynomials
K = interactive loci in the NK model
k = number of the experiments
M = phase numbers of the composition variables
m = parameters in very fast simulated re-annealing
N = loci length in the NK model
p = joint probability density function
$\tilde{p}$: = joint probability density function estimator
Q_x, Q_z = polynomial order of the composition and non-composition variables
S = information entropy
T = annealing temperature
T_0 = inital annealing temperature
U = information Energy
U_α = parameters in the RPV model
W_γ = parameters in the RPV model
$\vec{w}$ = non-composition variable vectors relative to the phase center
$\vec{X}$ = input variable vectors in a general gression neural network
$\vec{x}$ = composition variable vectors mole fraction
Y = output variables in a general gression neural network or merit function
$\tilde{Y}$ = predict value of output in a general gression neural network
$\vec{y}$ = composition variable vectors relative to the phase center
$\vec{z}$ = non-composition variable vectors mole fraction
α = composition phase index
γ = non-composition phase index
ξ_x, ξ_z = parameters in the RPV model
σ = smooth factor in a general regression neural network
σ_a = parameters in the NK model
σ_x, σ_z = parameters in the RPV model

Literature Cited

Bogarad, L. D. and M. W. Deem; "A Hierarchical Approach to Protein Molecular Evolution," *Proc. Natl. Acad. Sci.*, **96**, 2591–2595 (1999)

Bornscheuer, U. T.; "Directed Evolution of Enzymes," *Angew. Chem. Int. Ed.*, **37**, 3105–3108 (1998)

Chen, J., S. S. Jang, D. S. H. Wong and S. L. Yang; "Product and Process Development Using Artificial Neural-Network Model and Information Analysis," *AIChE J.*, **44**, 876–887 (1998)

Chen, S., C. F. N. Cowan and P. M. Grant; "Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks," *IEEE Trans. on Neural Net.*, **2**, 302–309 (1991)

Cole, B. M., K. D. Shimizu, C. A. Kreuger, J. P. A. Harrity, M. L. Snapper and A. H. Hoveyda; "Discovery of Chiral Catalysts through Ligand Diversity: Ti-Catalyzed Enantioselective Ad-

dition of TMSCN to Meso Epoxides," *Angew. Chem Int. Ed*., **35**, 1668–1671 (1996)

Cong, P., A. Dehestani, R. Doolen, D. M. Giaquinta, S. Guan, V. Markov, D. Poojary, K. Self, H. Turner and W. H. Weinberg; "Combinatorial Discovery of Oxidative Dehydrogenation Catalysts within the Mo-V-Nb-O System," *Proc. Natnl. Acad. Sci*., **96**, 11077–11080 (1999)

Danielson, E., M. Devenney, D. M. Giaquinta, J. H. Golden, R. C. Haushalter, E. W. McFarand, D. M. Poojary, C. M. Reaves, W. H. Weinberg and X. D. Wu; "X-Ray Powder Structure of Sr_2CeO_4: A New Luminescent Material Discovered by Combinatorial Chemistry," *Science*, **279**, 837–839 (1998)

Davis, M. E.; "Combinatorial Methods: How will They Integrate into Chemical Engineering?," *AIChE J*., **45**, 2270–2272 (1999)

Engstrom, J. R. and W. H. Weinberg; "Combinatorial Materials Science: Paradigm Shift in Materials Discovery and Optimization," *AIChE J*., **46**, 2–5 (2000)

Falcioni M., M. and W. Deem; "Library Design in Combinatorial Chemistry by Monte Carlo Methods," *Physical Review E*., **61**, 5948–5952 (2000)

Gordon, E. M.; Combinatorial and Molecular Diversity in Drug Discovery, pp. 345–417, Wiley, New York, USA (1998)

Gordon, E. M., M. A. Gallop and D. V. Patel; "Strategy and Tactics in Combinatorial Organic Synthesis. Applications to Drug Discovery," *Acc. Chem. Res*., **29**, 144–154 (1996)

Hanak, J. J.; "Multiple-Sample Concept in Materials Research: Synthesis, Compositional Analysis, and Testing of Entire Multicomponent Systems," *J. Mater. Sci*., **5**, 964–971 (1970)

Haykin, S.; Neural Networks-A Comprehensive Foundation, pp. 202–234, Prentice-Hall Inc., New Jersey, USA (1999)

Ingber, L.; "Very Fast Simulate Reanneling," *Math. Comput. Modeling*, **12**, 967–973 (1989)

Jandeleit, B., H. W. Turner, T. Uno, J. A. M. van Beck and W. H. Weinberg; "Combinatorial Methods in Catalysis," *CATTECH*, **2**, 101–123 (1998)

Kauffman, S. A.; The Origins of Order: Self Organization and Selection in Evolution, pp. 40–65, Oxford Univ. Press, New York, USA (1993)

Kauffman, S. A. and S. Levin; "Towards A General Theory of Adaptive Walks on Rugged Landscapes," *J. Theor. Biol*., **128**, 11–45 (1987)

Kirkpatrick, S., C. D. Gelatt, Jr. and M. P. Vecchi; "Optimization by Simulated Annealing," *Science*, **220**, 671–680 (1983)

Klein, J., C. W. Lehmann, H. W. Schmidt and W. F. Maier; "Combinatorial Material Libraries on the Microgram Scale with an Example of Hydrothermal Synthesis," *Angew. Chem. Int. Ed*., **37**, 3369–3372 (1998)

Lin, J. J. L., D. S. H. Wong and S. W. Yu.; "Optimal Multiloop Feedback Design Using Simulated Annealing and Neural Network," *AIChE J*., **43**, 430–434 (1995)

Linusson, A., J. Gottfries, F. Lindgren and S. Wold; "Statistical Molecular Design of Building Blocks for Combinatorial Chemistry," *J. Med. Chem*., **43**, 1320–1328 (2000)

Parzen, E.; "On Estimation of a Probability Density Function and Mode," *Ann. Math. Statist*., **33**, 1065–1076 (1962)

Perelson, A. S. and C. A. Macken; "Protein Evolution on Partially Correlated Landscapes," *Proc. Natl. Acad. Sci*., **92**, 9657–9661 (1995)

Pretel, E. J., P. A. Lopez, B. B. Susana and E. A. Brignole; "Computer-aid Molecular Design of Solvents for Separation Processes," *AIChE J*., **40**, 1349–1360 (1994)

Schlögl, R.; "Combinatorial Chemistry in Heterogeneous Catalysis: A New Scientific Approach or 'the King's New Clothes'?" *Angew Chem. Int. Ed*., **37**, 2333–2336 (1998)

Senkan S. M.; " High-Throughput Screening of Solid-State Catalyst Libraries," *Nature*, **394**, 350–353 (1998)

Shannon, C. E.; "A Mathematical Theory of Communication," *Bell Syst. Tech. J*., **27**, 379–423 (1948)

Specht., D. F.; "A General Regression Neural Network," *IEEE Trans. on Neural Net.*, **2**, 568–576 (1991)

Szostak, J, W.; "Introduction: Combinatorial Chemistry," *Chem. Rev*., **97**, 347–348 (1997)

Van Dover, R. B., L. F. Schneemeyer and R. M. Fleming; "Discovery of a Useful Thin-Film Dielectric Using a Composition-Spread Approach," *Nature*, **392**, 162–164 (1998)

Voigt, C. A., S. L. Mayo, F. H. Arnold and Z.-G. Wang; "Computational Method to Reduce Search Space for Directed Protein Evolution" *Proc. Natl. Acad. Sci*, **98**, 3778–3783 (2001)

Wilson, S. R. and A. W. Czarnik; Combinatorial Chemistry: Synthesis and Application, pp. 119–130, Wiley, New York, USA (1997)

Xiang, X. D., X. Sun, G. Briceno, Y. Lou, K. A. Wang, H. Chang, W. G. Wallace-Freedman, S. W. Chen and P. G. Schultz; "A Combinatorial Approach to Materials Discovery," *Science*, **268**, 1738–1740 (1995)

You, L. and F. H. Arnold; "Directed Evolution of Subtilisin E in *Bacillus subtilis* to Enhance Total Activity in Aqueous Dimethylformamide," *Protein Eng*., **9**, 77–83 (1996)