Journal of Natural Sciences
Pai Chai University, Korea
Vol. 15, No. 1 : 11-34, 2004

Multiensemble Sampling 방법의 속성에 대한 연구

한 규 광
대전광역시 서구 도마2동 배재대학교 물리학과

A study on the properties of the Multiensemble Sampling method

Kyu-Kwang Han
Department of Physics, PaiChai University
439-6, Doma-2-dong, Taejon 302-735, Korea

요 약

전산분자시늉(computational molecular simulation)을 이용하여 생물분자와 같이 다양하고 복잡한 분자계들을 연구하는데 있어서, 그 성과는 표본추출 알고리즘의 배열공간(configuration space) 탐사능력에 의해 결정된다고 말해도 과언이 아니다. 본 연구에서는 기존의 방법들이 안고 있는 문제점들을 제거 또는 완화시킬 수 있는 유력한 방법 중 하나인 Multiensemble Sampling (MES) 방법의 실용적 속성에 대한 체계적인 연구로부터, 분자배열공간상에서 떨어져있는 영역들을 탐사하는데 MES를 효과적으로 적용할 수 있는 체계적이면서도 일반적인 실행 방법론을 도출해 내기 위한 작업을 수행하였다. 이 작업에서는 좀 더 일반화시킨 MES의 무게함수의 꼴을 이용하여, '물 속에서의 공동 형성'을 몬테카를로(MC)로 수행하였다. 여러 경우에 대해 시늉내기 매개변수들과 계산의 효율성과의 상관관계를 조사하여 그 결과로부터 효율성을 극대화할 수 있는 실행 체계를 제시할 수 있었다. 그 체계에 따라 반지름이 0에서부터 5.64A에 이르는 여러 크기의 공동을 갖는 계들을 한 번의 시늉으로 탐사할 수 있었다.

Abstract

It is no exaggeration to say that the productivity of a research using computer simulations on complex molecular systems like biomolecules depends on the ability of the sampling algorithm to

explore the relevant parts of configuration space. In this study, we investigate the properties of the multiensemble sampling (MES) which is one of the solutions that surmount limitations of conventional sampling algorithms. Works for finding out practical systematic ways of using the MES efficiently to explore distantly separated regions in configuration space are performed. In this work, the more generalized form of weighting function for MES is used and `cavity formation in water' is simulated using Monte Carlo. Investigating the correlation of simulation parameters and the efficiency of the method, we propose a practical way of maximizing the power of the MES. We applied the way to `cavity formation in water' and were able to explore the parts of configuration space relevant to cavities of radius from 0 to 5.6A in a single simulation.

I. 서 론

생명과학의 발달은 놀랍도록 빠르게 진행되고 있으며, 그 연구결과의 응용분야가 매우 넓어 미래의 인류생활에 커다란 영향을 미치게 될 것이 확실하다. 21세기의 생명과학은 물리학과 화학의 지식을 기초로 한 생명현상에 대한 종합적인 탐구와 함께 그 동안 축적되어 온 과학적 지식과 방법을 실생활에 응용하는 학문으로 발전하게 될 것이다. 현재, 이러한 학제간의 협동 연구는 아래의 예들과 같이 전산을 매개로 하여 이루어지고 있다.

유전자 분석 등의 방법을 통하여 유전자 및 단백질의 아미노산 서열은 이미 수만개나 발표되어 있다. 그런데, 대부분의 경우에 생체 물질의 결정을 만들기가 쉽지 않아, 실험적으로 그 구조를 밝히는 데에 어려움이 있다. 아미노산 서열(즉, 단백질의 2차구조)은 알려져 있으나, 그 단백질의 3차원 원자좌표가 결정되지 않았다면, 이러한 단백질의 구조를 Modeller 등과 같은 "homology" modeling 프로그램을 이용하여 예측할 수 있다.[1,2] 이러한 기법을 이용하여, 생체 거대 분자의 3차원 구조를 모르는 경우라도, 모델링 결과를 이용하여 분자 시뮬레이션을 수행하여 쓸만한 결론을 도출할 수 있다. 단백질, 핵산, 지질, 당 등의 생체 거대분자의 분자 역학 모형을 작성하고 이것을 물분자 등으로 용매화 시켜서 생체 고분자의 분자적 거동을 전산시늉(computer simulation)하여 고찰 할 수도 있다.[3,4] 효소와 기질, 수용체와 전달체, 핵산과 결합단백질, 생체막 등에서 일어나는 생명현상을 분자 사이의 상호작용으로 분석함으로

써 분자생물학적 과정에 대한 논의를 원자 단위로 구체화한다.

전산분자시늉(computational molecular simulation)을 이용하여 생물분자와 같이 다양하고 복잡한 분자계들을 연구하는데 있어서, 그 성과는 분자상호작용 모형의 정확성과 표본추출 알고리즘의 배열공간(configuration space)탐사능력에 의해 결정된다고 말해도 과언이 아니다. 이들 중에 후자의 향상이 선행되어 한다. 왜냐하면, 후자의 문제가 먼저 해결되어야 전자의 문제점을 평가할 수 있기 때문이다. 몬테카를로(MC)와 분자동력학(MD) 등을 이용한 전산분자시늉 실험이 보편화됨에 따라 시늉방법의 효율성을 개선하는 일은 중요한 과제중 하나가 되었다. 최근의 컴퓨터의 발전 덕분에 생물분자와 같은 크고 복잡한 계들에 대한 시늉실험연구들이 가능하게 되면서, 그 중요성은 더해지고 있다. 특히, 효과적인 배열공간탐사를 위한 표본추출방법의 개선이 그렇다. 예를 들면, 분자시늉실험을 통한 자유에너지 계산은 열역학적, 화학적, 생화학적 현상들의 에너지학적 질문들에 대해 설명할 수 있는 유력한 수단이며 그 적용 가능 분야는 solvation, association, macromolecular stability, enzyme catalysis, enzyme-ligand binding energies site-directed mutagenesis, host-guest studies, conformational changes 등 열거하기 벅찰 정도로 다양하다.[5-14] 그런데 이렇게 다양한 분야에 자유에너지 계산을 적용하는 데에 있어서 걸림돌이 있다. 즉, 현재 보편적으로 쓰이고 있는 방법들은 탐사되어져야 하는 계의 배열공간을 제대로 탐사하지 못하기 때문에, 적합한 구조들을 표본 추출하는 데에 한계가 있다. 이런 장애는 계의 크기와 구조가 커지고 복잡해질수록 심각해진다.

배열공간탐사능력 향상에 이용될 수 있는 알고리즘들로는 umbrella sampling (US)[15-18], mutiensemble sampling (MES)[19-21], multicanonical algorithm (MCA)[22,23], entropic sampling[24,25], simulated tempering (ST)[26,28], replica_exchange method (REM)[28-30] 등의 여러 방법들이 있다. 이 방법들은 적절한 무게함수가 주어진다는 가정 하에서 우수하다. 그러나 MES와 REM를 제외한 대부분의 방법들에서는 무게함수의 꼴을 미리 알 수가 없기 때문에 그 효율성이 떨어진다. 이런 어려움을 극복하기 위하여 최근에 개발된 방법들이 있으나, 전체적 최소 에너지 (global-minimum energy) 값이 주어질 때만 무게함수를 알 수 있는 것이어서, 전체적 최소 에너지 측정에 따르는 어려움이 여전히 있다. REM에서는 상호작용하지 않는 여러 개의 계(replica)들이 동시에 독립적으로 시늉되어지면서 약간의 매 단계마다 정해진 확률에 따라 replica 쌍들이 맞바꾸어진다. REM의 무게함수는 볼츠만 인자의 곱이다. 그러나 REM 역시 계산의 어려움이 있다. 계의 자유도가 늘어남에 따라, 즉, 계가 복잡해짐에 따라 소요되

는 replica의 수가 엄청나게 늘어나기 때문이다. 하나의 replica만을 시늉하는 US, MCA or ST에 비해 엄청난 계산이 필요하게 된다. 반면, MES는 하나의 replica만을 시늉한다는 것이 US, MCA, ST와 같으나, 볼츠만 인자들의 조합으로 이루어진 만능의(universal) 무게함수 꼴이 주어진다는 것이 다르다. MES에서 시늉되는 replica는 여러 개의 계들의 조합으로 이루어진 것이어서, MCA, ST의 장점과 REM의 장점을 고루 갖추었고 할 수 있다. MES는 REM 보다 먼저 제안되었으나 그 적용 예가 적어서 성능의 우수성이 아직 잘 알려지지 않은데다가, 속성에 대한 연구부족 등으로 널리 쓰이지 않고 있다.

우리는 MES의 실용적 속성에 대해 연구하고자 한다. 연구의 목적은 MES의 체계적이면서도 일반적인 실행 방법을 도출해 내어 배열 공간을 효율적으로 탐사할 수 있는 고유기술을 개발함에 있다. 국외에서는 효율적인 배열공간탐사 알고리즘들의 개발과 이용이 보편화되어 가고 있는데 비해, 국내에서는 아직 그렇지 못한 상태이다. MES를 택한 이유는 그 알고리즘이 본 과제 책임자가 제안한 것으로서 독창성이 있을 뿐 아니라, 최근 발표된 이온수화에 적용한 연구 결과는 그 우수성을 입증하여주고 있기 때문이다. 본 연구는 MES를 패키지화하기 위한 전초적인 작업이기도 하다.

이제까지의 MES를 이용한 연구에서는 기존의 방법들보다 더 효율적임을 보이는 데에 그쳤다. 본 연구에서는 분자배열공간상에서 떨어져있는 두 영역을 탐사하는데 MES를 효과적으로 적용할 수 있는 체계적이면서도 일반적인 실행 방법론을 도출해 내기 위한 작업을 수행한다. 이 작업에서는 좀 더 일반화시킨 MES의 무게함수의 꼴을 이용하며, '물 속에서 여러 크기의 공동 형성'을 몬테카를로(MC)로 수행하여, 시늉내기 매개변수들과 계산의 효율성과의 상관관계를 조사하고 그 결과로부터 효율성을 극대화할 수 있는 체계적인 실행방법을 제시한다.

II. 본 론

1. 이 론

가. Boltzmann sampling 체계와 non-Boltzmann sampling 체계

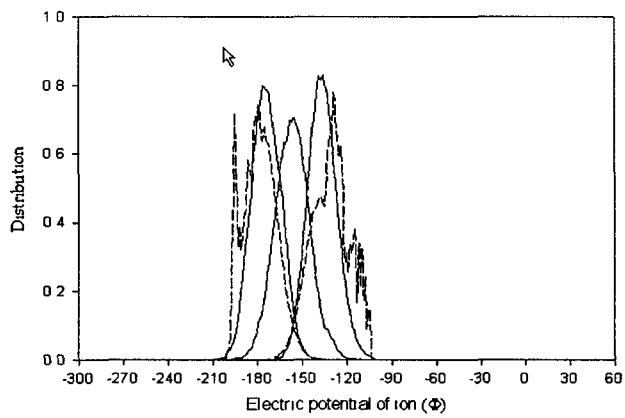
먼저 배열공간탐색에 있어서 통상적인 Boltzmann sampling 체계의 문제점을 자유에너지 계

산에 적용하여 이론적으로 재조명해보자.

계 l과 m의 자유에너지 차이 ΔF_{ml}의 계산에 있어서, 기존 방법들은

$$e^{-\beta\Delta F_{ml}} = \frac{Z_m}{Z_l} = \frac{\int e^{-E_m/kT} d\Omega}{\int e^{-E_l/kT} d\Omega}$$
$$= \frac{\int e^{(E_l-E_m)/kT} e^{-E_l/kT} d\Omega}{\int e^{-E_l/kT} d\Omega} = \langle e^{(E_l-E_m)/kT} \rangle_l \qquad \text{(II-1)}$$
$$= \frac{\int e^{-E_m/kT} d\Omega}{\int e^{(E_m-E_l)/kT} e^{-E_m/kT} d\Omega} = \langle e^{(E_l-E_m)/kT} \rangle_m$$

와 같이 계 l 또는 m 에 대한 Boltzmann sampling으로부터의 계산에 근본을 두고 있다. 여기서, $\langle\rangle_l$과 $\langle\rangle_m$은 각각 E_l과 E_m계의 앙상블 (또는 시간) 평균을 뜻한다. 수식 II-1에 의한 계산은 두 계가 아주 유사하지 않으면, E_l 계에 대해서만 표본추출(sampling)된 분자배열들이 배열공간(configuration space)상에서 차지하는 영역이 E_m 계의 중요한 영역을 덮지 못하기 때문에, 다시 말해서 두 계의 공유 영역이 충분히 크지 않기 때문에, 정확한 값을 얻기가 어렵다. <그림 II-1>은 이것은 잘 보여주고 있다. <그림 II-1>은 수식 II-1을 이온의 전하량 변화에 따는 수화자유에너지 차이 계산에 적용한 예를 보여주는 것이다. 가운데 실선이 실제로 표본추출된 배열들의 분포이고, 좌우의 점선들은 그 분포로부터 뽑아낸 두 이웃 전하량들에 대한 분포들인데 그들의 참 Boltzmann 분포(양쪽 실선)와 많이 다름을 알 수 있다.



<그림 II-1> Ilustration of the limitation in Boltzmann sampling scheme

이런 단점을 개선하기 위해서는 이웃하는 계들의 중요한 영역을 덮을 수 있는 일반적인 무게함수 W를 이용하는 non-Boltzmann sampling 체계로 추출하여야 한다. 그러면 자유에너지 다음의 식으로 계산할 수 있다.

$$e^{-\Delta F_{ml}/kT} = \frac{\int e^{-E_m/kT} W^{-1} W d\Omega / \int W d\Omega}{\int e^{-E_l/kT} W^{-1} W d\Omega / \int W d} = \frac{\langle e^{-E_m/kT} W^{-1} \rangle_W}{\langle e^{-E_l/kT} W^{-1} \rangle_W} \qquad \text{(II-2)}$$

서론에서 언급한 umbrella sampling (US), mutiensemble sampling (MES), multicanonical algorithm (MCA), entropic sampling, simulated tempering (ST), replica_exchange method (REM) 들은 모두 non-Boltzmann sampling 체계에 근거를 두고 있다. 두 계의 Boltzmann 분포를 모두 덮을 수 있는 분포함수 W를 택한다면, 수식 II-2는 수식 II-1보다 정확한 결과를 주게 된다. 그런데, 대부분의 non-Boltzmann sampling 방법에서는 W의 만능적인 함수꼴(universal functional form)이 주어지지 않았기 때문에 두 관심 계에 대해 미리 잘 알고 있어야만 하고 그 효율성도 불확실하다.

나. Multiensemble sampling (MES)

MES에서는 W의 꼴이

$$W \propto [e^{2(C_l - E_l)/kT} + e^{2(C_m - E_m)/kT}]^{1/2} \qquad \text{(II-3)}$$

와 같이 두 계에 대한 Boltzmann 분포의 중첩으로 주어진데, 이 꼴은 수식 II-2의 이론적으로 예상되는 오차에 대한 범함수적 최소화(functional minimization)로부터 얻어진 것이다.[19] 여기서 C들은 임의의 상수들인데, $C_m - C_l = \Delta F_{ml}$일 때에 이론적으로 예상되는 오차가 최소이다. $C_{ml} = \pm\infty$ 일 때에 수식 II-1로 환원됨을 주목하기 바란다. 이것은 C_{ml}를 어떻게 잡더라도 수식 II-2에 의한 계산이 수식 II-1 에 의한 것보다는 더 정확하다는 것을 의미한다. 실제 계산에 있어서는 처음에 $\Delta C_{ml} = C_m - C_l = 0$으로 놓고 시작하여 적당한 시늉내기 간격 동안에 계산되어지는 ΔF_{ml}의 값으로 대치하는 것을 반복한다. 이런 반복에 의한 ΔC_{ml}의 ΔF_{ml} 참값으로의 수렴은 지수함수적으로 빠르다.[19-21] <그림 II-2>은 L-J fluid와 inverse-twelve fluid에 대한 MES 결과로서, 10^6 MC step 동안 추출된 분자배열들의 에너지에 대한 분포를 그린 것이다 (실선). 이 그림은 두 계의 분자배열 분포의 겹침이 극단적으로 나쁘지 않으면,

계들의 앙상블을 동시에 제대로 추출할 수 있음을 보여주고 있다. 수식 II-1을 사용한다면 점선의 분포들중 하나만을 추출하게 되므로 이 두 계사이의 자유에너지 차이를 계산하는 것은 불가능하다. 그래서 기존의 방법에서는, 두 계 사이에 많은 가상계들(virtual systems)을 두어 가까운 두 계 사이의 자유에너지 차이를 계산해서 합하는, free energy perturbation 방법을 쓰게 되는 것이다. 수식 II-3은 여러 계에 대한 Boltzmann 분포의 중첩을 추출할 수 있도록

$$W \propto \left[\sum_{i=1}^{n_l} e^{2(C_i - E_i)/kT} \right]^{1/2} \qquad \text{(II-4)}$$

같이 일반화할 수 있으며[20], 계 l 과 계 m 사이의 자유에너지 차이는

$$e^{-\Delta F_{ml}/kT} = e^{-\Delta C_{ml}/kT} \frac{\langle f_m \rangle_W}{\langle f_l \rangle_W} \qquad \text{(II-5)}$$

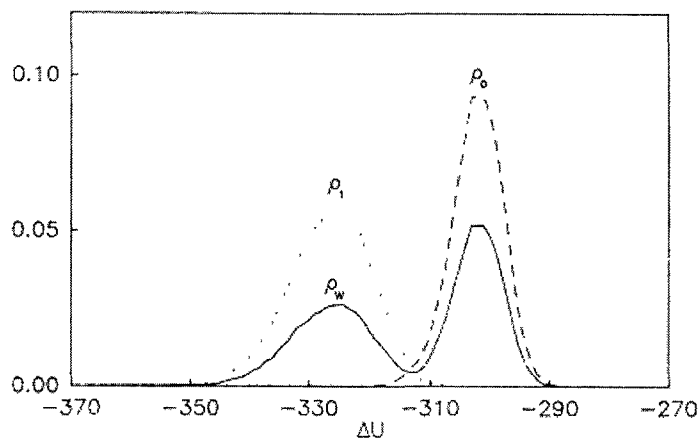


Fig. 1. Plots of densities of configurations versus ΔU.

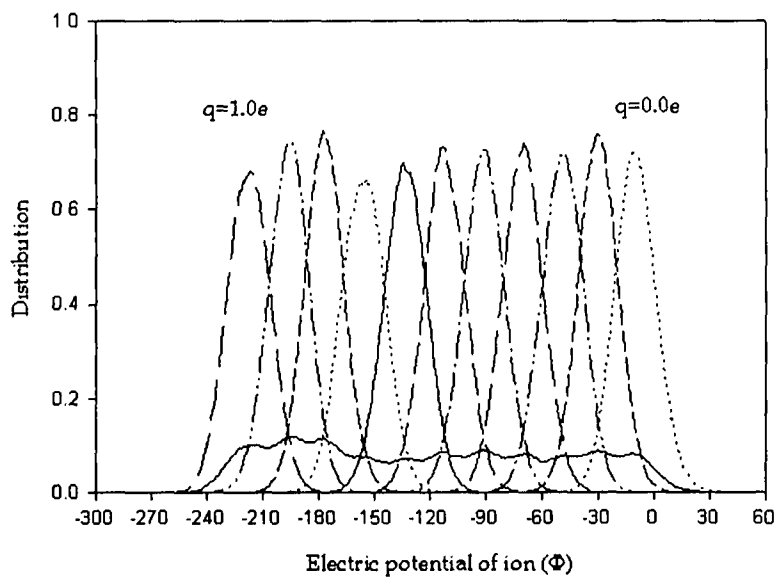
<그림 II-2> Application of MES to L-J fluid and inverse-twelve fluid

을 계산하여 얻을 수 있다. 여기서

$$f_m = \{1 + \sum_{l \neq m} \exp[-2(\Delta C_{ml} - \Delta E_{ml})/kT]\}^{-1/2} \qquad \text{(II-6)}$$

이다.

최근에 우리는 물에서 용해되어 있는 Ca^+ 이온과 Ca 원자의 자유에너지 차이를 MES로 계산였다.[21] 이 경우에는 두 계의 물분자 배열이 달라서 그 분포가 전혀 겹치지 않는다. 하지만, 9개의 가상의 중간계를 포함한 11개 계들에 대한 Boltzmann 분포의 중첩을 한번의 시늉내기로 추출함으로써 자유에너지 차이를 정확하게 계산할 수 있었다. 또한, 기존의 섭동방법과 성능을 비교하기 위하여 SP도 적용해 보았다. 위의 11개 계들을 각각 시늉내기하여 얻은 결과는 MES보다 오차가 훨씬 컸으며 역 방향 계산에서는 이력현상이 관찰되었다. 우리는 두 방법의 신뢰성을 비교하기 위하여 두 방법에 의해 추출된 분포로부터 계산에 포함된 계들의 Boltzmann 분포를 뽑아내어 보았다. <그림 II-3> 는 MES의 결과로서, 물 분자배열에 의한 Ca 이온의 전위 분포를 그린 것인데 (실선), 각 계들에 대한 분자배열 분포가 공평하게 추출되었음을 알 수 있다. 점선의 분포들은 추출한 실선 분포로부터 뽑아낸 각 계들의 Boltzmann 분포들인데 기존의 Boltzmann sampling 결과와 일치한다.



<그림 II-3> Application of MES to the ionization

이제까지의 MES를 이용한 연구에서는 기존의 방법들보다 더 효율적임을 보이는 데에 그쳤다. 본 연구에서는 분자배열공간상에서 떨어져있는 두 영역을 탐사하는데 MES를 효과적으로 적용할 수 있는 체계적이면서도 일반적인 실행 방법론을 도출해 내기 위한 작업을 수

행한다. 이 작업에서는 좀 더 일반화시킨 MES의 무게함수의 꼴을 이용하며, '물 속에서의 여러 크기의 공동 형성'을 몬테카를로(MC)로 수행하여, 시늉내기 매개변수들과 계산의 효율성과의 상관관계를 조사하고 그 결과로부터 효율성을 극대화할 수 있는 체계적인 실행방법을 제시한다.

한편, 수식 II-4은 다음과 같이 좀 더 일반화할 수 있다.

$$W \propto \left[\sum_{i=1}^{n} e^{p(C_i - E_i)/kT} \right]^{1/p} \qquad \text{(II-7)}$$

여기서 p는 표본 추출되는 분자배열의 분포의 속성을 결정하게 되므로 속성변수라고 할 수 있겠다. 이론적으로는 p=2 일 때에 예상되는 오차가 최소이다. 본 연구에서는 p 값의 변화에 따른 계산의 정확도와 수렴도의 변화를 조사하여 MES의 효율성 향상 가능성을 조사하는 작업을 병행한다.

2. 시늉과 결과

가. 물 속의 부드러운 공동 형성에 대한 몬테카를로 전산시늉

온도 298.15 K, 밀도 1g/cm^{3}인 물 속에서 여러 크기의 공동들에 대한 배열공간탐사를 수식 II-7의 p = 0.5, 1, 2, 4인 무게함수들에 따라 몬테카를로 시늉하여, 수식 II-5로 표현되는 동공 형성을 위한 자유에너지를 계산하였다. 시늉 계는 한 변의 길이가 15.52 A인 정육면체 상자 안에 있는 125 TIP4P 물 분자들과 하나의 공동으로 이루어진다. 주기적 경계조건 (periodic boundary conditio) 아래에서 분자들 사이의 상호작용들이 계산되었다. 물 분자들 사이의 상호작용은 7.75 A에서 끊어 버렸다. 공동과 그 주위 물분자들에 대한 통계를 강화하기 위해 preferential sampling 알고리즘을 사용하였다. 본 연구에서는 물분자의 움직임을 시도하는 확률을 1/r에 비례하게 잡았는데, 여기서 r은 공동과 물분자 사이의 거리이다. 물 속에서의 부드러운 공동 형성을 위하여 다음과 같이 r^{-12}의 반발 포텐셜 함수를 사용하였다.

$$\Phi_{cw} = \frac{(A_c A_w)^{1/2}}{r^{12}} \qquad \text{(II-8)}$$

여기서 A_c와 A_w는 다음과 같이 정의되는 통상적인 Lennard-Jones 반발 매개 변수이다.

$$A = 4u\sigma^{12} \qquad \text{(II-9)}$$

여기서 u는 포텐셜 우물의 깊이이고, σ는 단단한 핵심 반지름인데, u_c = u_w =0.155 kcal/mol, $\sigma_w = 3.1536$A으로 하고 σ_c 는 0A에서 10A까지 0.5A 간격으로 21개의 값들로 잡았다. 공동의 크기는, $\Phi_{cw} = kT$가 되게끔, 다음과 같이 열적 반지름(thermal radius) r_{th}를 정의하여 나타낼 수 있다.

<table II-1>에 우리의 탐사 대상이 된 계와 그 공동의 크기를 나열하였다.

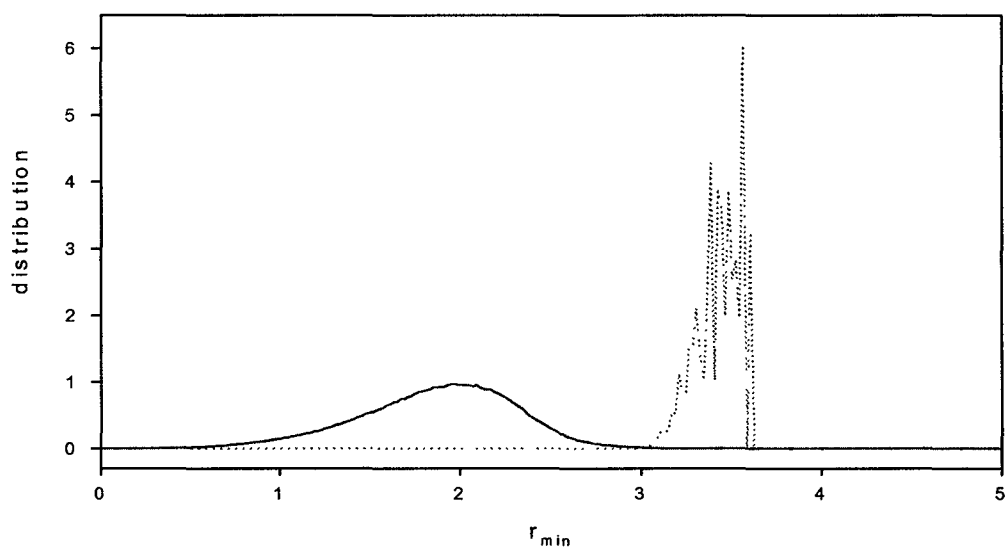
$$r_{th} = \left[\frac{(A_c A_w)^{1/2}}{kT} \right] \qquad \text{(II-10)}$$

<table II-1> Simulated systems

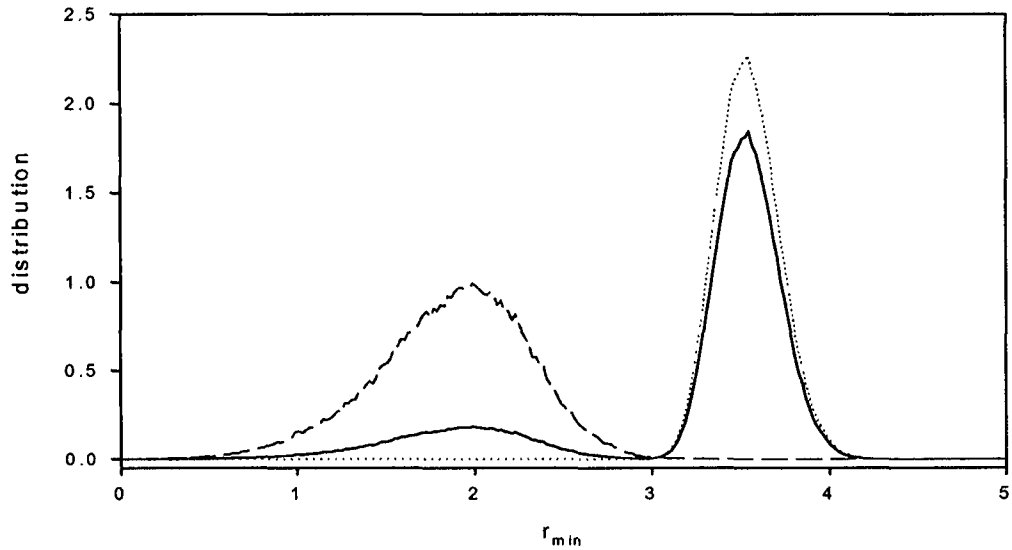
System	σ_c	r_{th}
S0	0	0
S1	0.5	1.26
S2	1.0	1.78
S3	1.5	2.18
S4	2.0	2.52
S5	2.5	2.82
S6	3.0	3.09
S7	3.5	3.33
S8	4.0	3.57
S9	4.5	3.78
S10	5.0	3.99
S11	5.5	4.18
S12	6.0	4.37
S13	6.5	4.54
S14	7.0	4.72
S15	7.5	4.88
S16	8.0	5.04
S17	8.5	5.20
S18	9.0	5.35
S19	9.5	5.49
S20	10.0	5.64

나. 두 계에 만의 배열공간 탐사

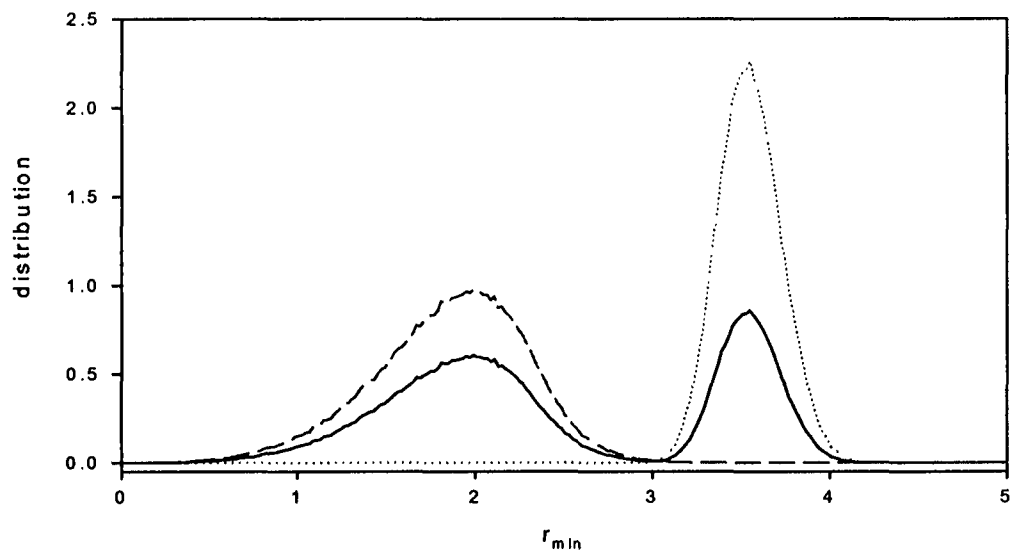
먼저, 두 계에 만의 배열공간을 탐사하는데 있어서 그 두 계의 다른 정도와 수식 II-7의 p 값에 따른 MES의 능력을 조사 하여보았다. 이를 위하여 σ_c 가 1.0A, 2.0A, 3.0A 그리고 4.0A인 공동형성 자유에너지를 계산하였다. 모든 경우에 대해서, 앞서 이론에서 언급하였듯이, 처음에 $\Delta C_{ml} = C_m - C_l = 0$으로 놓고 시작하여 적당한 시늉내기 간격 동안에 계산되어지는 자유에너지 ΔF_{ml}의 값으로 대치하는 것을 반복하였다. 한번의 시늉에 있어서 2×10^6에서 10^7 배열들이 표본 추출되었으며 각 경우에 대해 2-3회 반복시행 하였다. 우리는 p = 0.5, 1.0, 2.0, 4.0의 값들을 사용하여 대체로 비슷한 결과를 얻었다. <그림 II-4> - <그림 II-10>은 p=2를 사용하였을 때 표본 추출된 배열들의 공동과 가장 가까운 물분자사이의 거리에 대한 분포들을 그린 것이다.



<그림 II-4> Distribution of the distance between the cavity and its nearest water molecule : results of the first simulation using p=2 for S0 and S8 ($\Delta C_{80}^{(1)} = 0$)



<그림 II-5> Distribution of the distance between the cavity and its nearest water molecule results of the second simulation using p=2 for S0 and S8 ($\Delta C_{80}^{(2)} = \Delta F_{80}^{(1)}$)

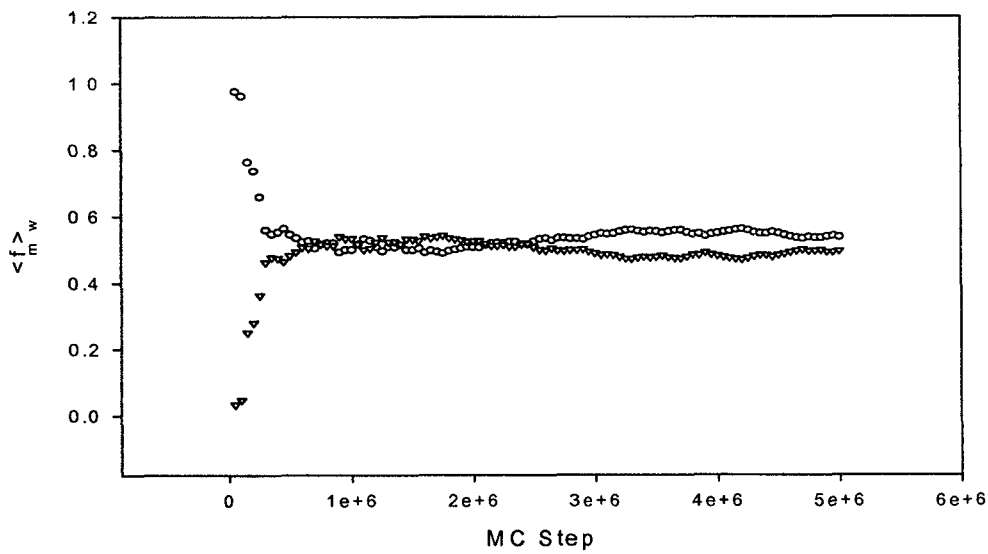


<그림 II-6> Distribution of the distance between the cavity and its nearest water molecule results of the third simulation using p=2 for S0 and S8 ($\Delta C_{80}^{(3)} = \Delta F_{80}^{(2)}$)

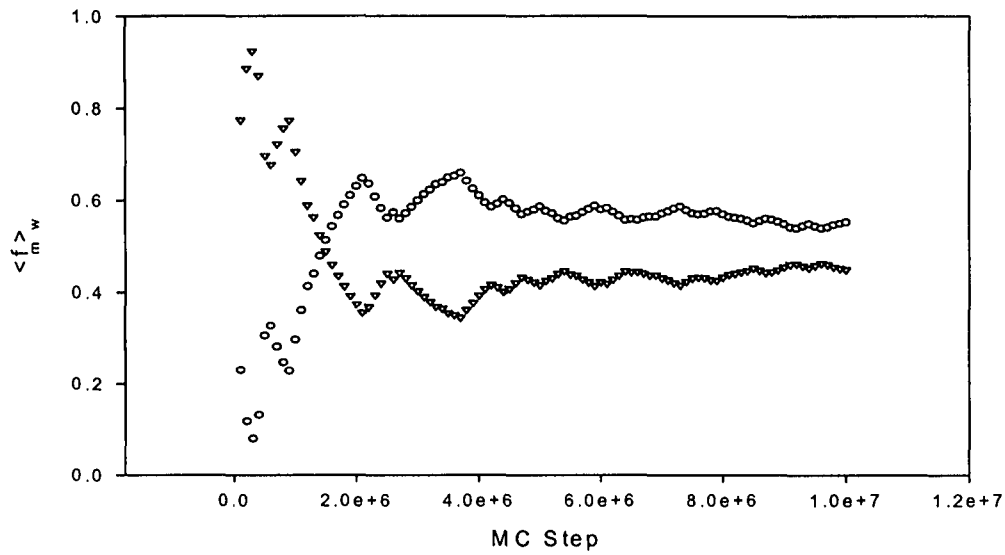
그림들에서 실선은 표본추출된 배열들의 분포 ρ 이고 점선들은 계들의 개별적 분포들 ρ_l 인데 다음의 식으로 계산되어진 것이다.

$$\rho_l = \rho \frac{f_l}{\langle f_l \rangle} \qquad \text{(II-11)}$$

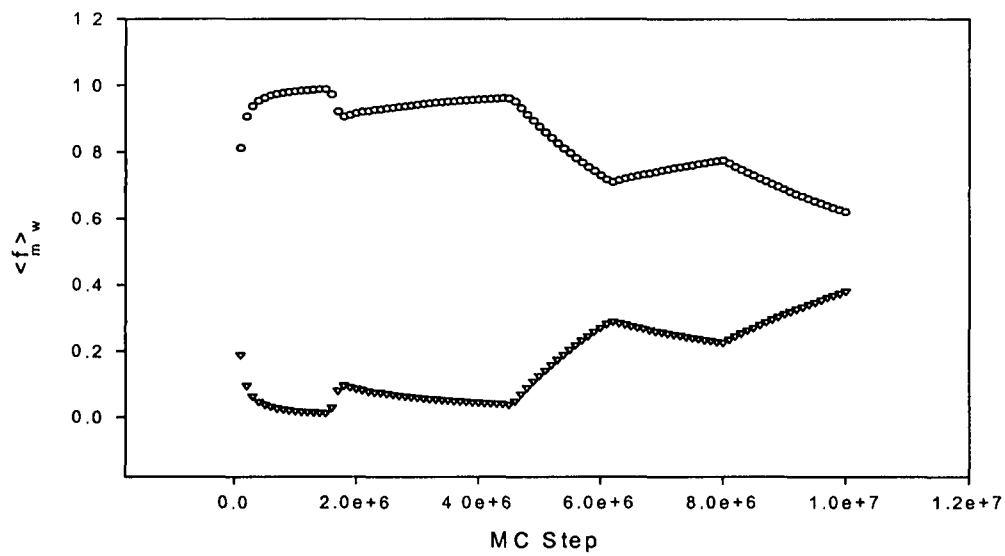
<그림 II-4>와 <그림 II-6>는 공동이 없는 계 S0와 열적 반지름이 3.57A 인 공동을 가진 계 S8에 대한 것을 시늉 순으로 그린 것들이다. 이 그림들은, 시늉이 반복될수록 두 계의 배열공간을 공평하게 탐사하게 됨을 보여주고 있다. <그림 II-7> - <그림 II-9>은 마지막 시늉들에서 수식 II-6의해 계산된 f_m에 대한 누적 평균, $\langle f_m \rangle_w$을 몬테카를로 걸음 수에 대한 함수로 그린 것이다. 이 그림들로부터 두 계가 공통으로 차지하는 배열공간이 작을수록 그 수렴이 느려짐을 알 수 있다. 왜냐하면, 수식 II-5를 보면 알 수 있듯이, 모든 m에 대해 $\langle f_m \rangle_w$가 같은 값을 갖게 된다면 자유에너지 계산이 참값으로의 수렴되었음을 뜻하는 것이기 때문이다.



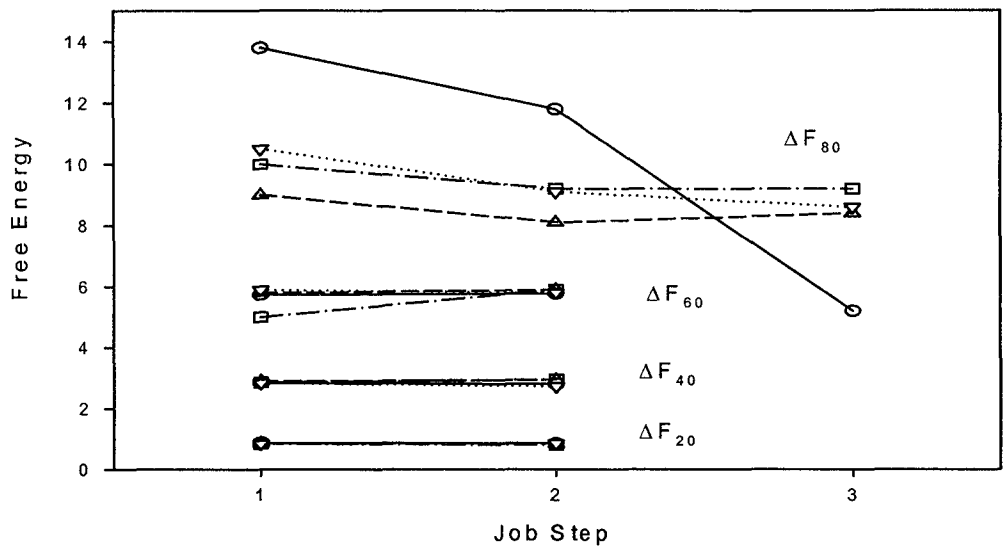
<그림 II-7> $\langle f_0 \rangle$ and $\langle f_4 \rangle$ vs MC step in the simulation for SO and S4



<그림 II-8> $\langle f_0 \rangle$ and $\langle f_6 \rangle$ vs MC step in the simulation for SO and S6



<그림 II-9> $\langle f_0 \rangle$ and $\langle f_8 \rangle$ vs MC step in the simulation for SO and S8



<그림 II-10> free energy vs job step (○: p=0.5, □: p=1, △: p=2, ▽: p=4)

<그림 II-10>은 사용된 p 값들에 대해, 각 시늉 단계마다 계산된 공동형성 자유에너지 값들을 보여준 것이다. 이 그림은 배영공간상에서 멀리 떨어져 있는 두 계를 탐사하는 데에 있어서, p 값이 작을수록 불리하다는 것을 의미한다.

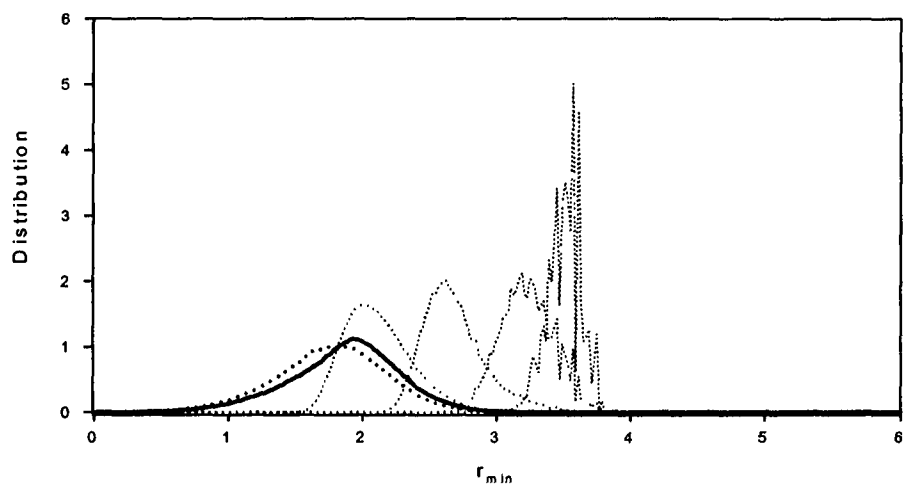
다. 세 개 이상의 계들의 배열공간탐사

배열공간상에서 멀리 떨어져있는 두 계를 탐사해야 할 경우, 그 사이에 다리 역할을 하는 중간 계들을 두어야 한다. (이렇게 하는 것을 통상, 창들을 연다고 한다.) 앞 절에서도 그랬듯이, 처음에 $\Delta C_{ml} = C_m - C_l = 0$으로 놓고 시작하여 적당한 시늉내기 간격 동안에 계산되어지는 자유에너지 ΔF_{ml}의 값으로 대치하는 것을 반복하는 실험을 p = 1 을 사용하여 두 가지 하였다. 하나는 S0, S2, S4, S6 그리고 S8 5개의 계들에 대한 것이고, 다른 하나는 S0, S1, S3, S4, S5, S6, S7, S8 9개의 계들에 대한 것이다. 각 실험에서 3번의 시늉이 반복되었다. 각 시늉마다 10^7 배열들이 표본 추출되었다. <table II-2>에 각 시늉에서 계산된 자유에너지 값들이 나열되어 있다. <그림 II-11> - <그림 II-13>은 첫 번째 실험에서 순차적으로 얻은 배열분포들이고, <그림 II-14> - <그림 II-16>은 두 번째 실험에서 얻은 것들이다. 이 그림들로부터 창을 많이 열수록 MES의 탐사능력이 향상됨을 알 수 있다.

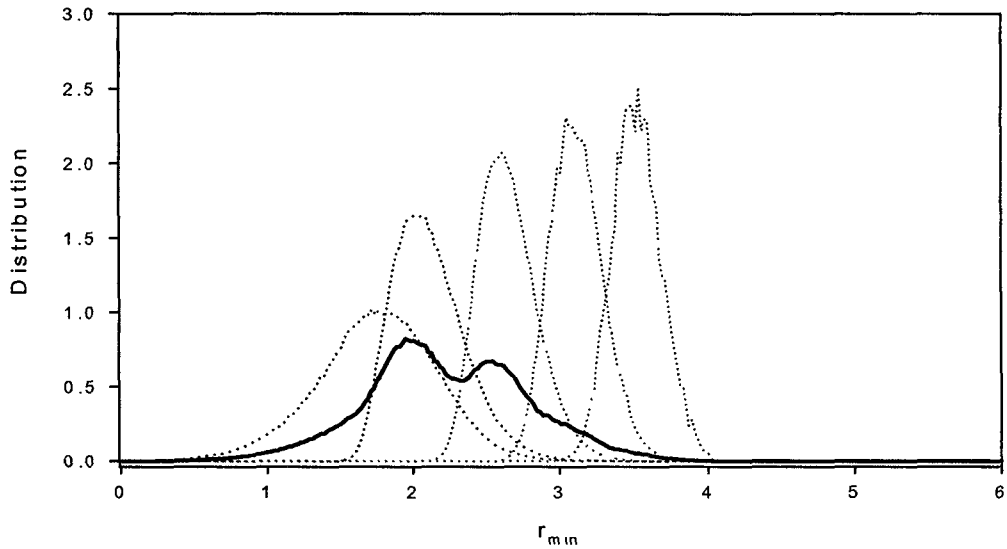
<table II-2> Free energies calculated in individual simulations

	실 험 1			실 험 2		
r_{th}	시늉1	시늉2	시늉3	시늉1	시늉2	시늉3
1.26				0.119	0.115	0.116
1.78	0.575	0.592	0.625	0.621	0.626	0.621
2.18				1.570	1.572	1.534
2.52	2.601	2.732	2.917	2.843	2.788	2.705
2.82				4.345	4.227	4.115
3.09	4.880	5.778	6.044	5.954	5.906	5.738
3.33				7.875	7.796	7.637
3.57	7.674	9.473	9.374	10.624	9.906	9.913

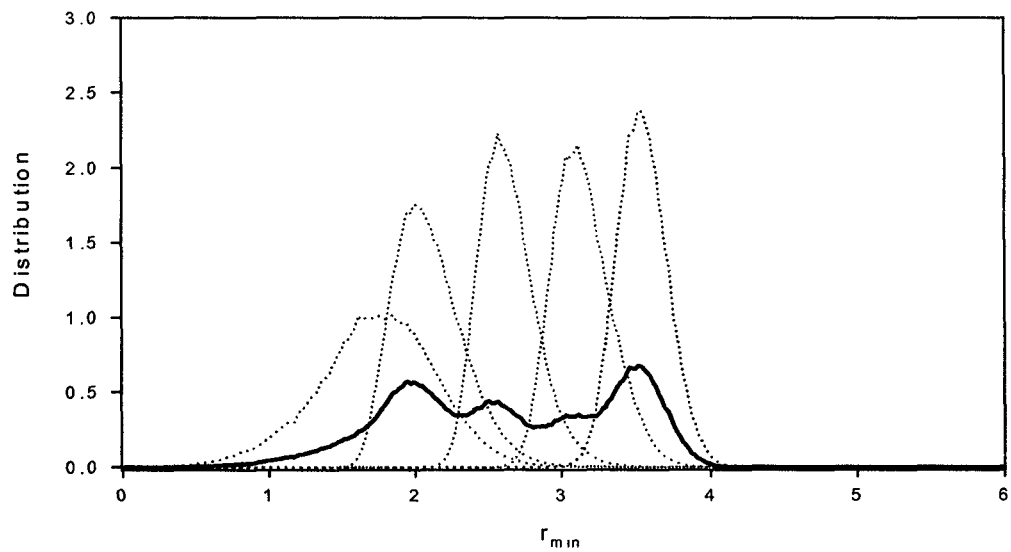
그런데 배열공간상에서 탐사해야할 영역들이 너무 넓은 경우에는 이와 같이 매개변수 ΔC_{ml} 을 한꺼번에 조정하는 것은 거의 불가능하기 때문에 Sm과 Sm+1 쌍들에 대한 예비시늉을 통하여 $\Delta C_{m,m+1}$ 을 개별적으로 조정하는 것이 효율적이다. 그러나 이것을 실행하기 전에 얼마 많은 창을 열 것인지가 결정되어야 한다. 계산의 효율성을 높이기 위해서는 적당한 수의 창들이 열려야 한다. 너무 많은 수의 창을 열면, $\Delta C_{m,m+1}$ 값을 개별적으로 조정하는 예비시늉의 수가 쓸데없이 너무 많아질 수 있고,



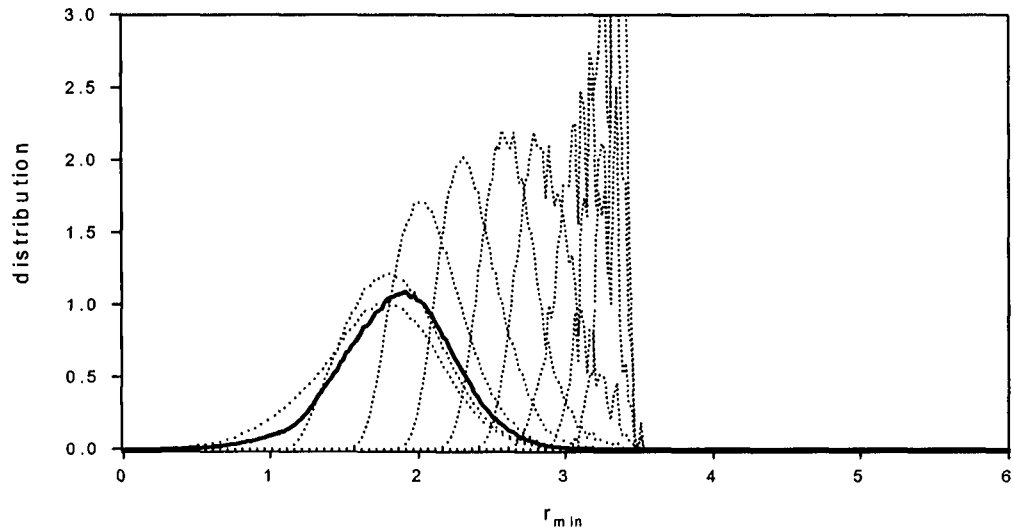
<그림 II-11> Distribution of the distance between the cavity and its nearest water molecule : results of the first simulation using p=2 for S0, S2, S4, S6 and S8 ($\Delta C_{m0}^{(1)} = 0$, $m = 2, 4, 6, 8$)



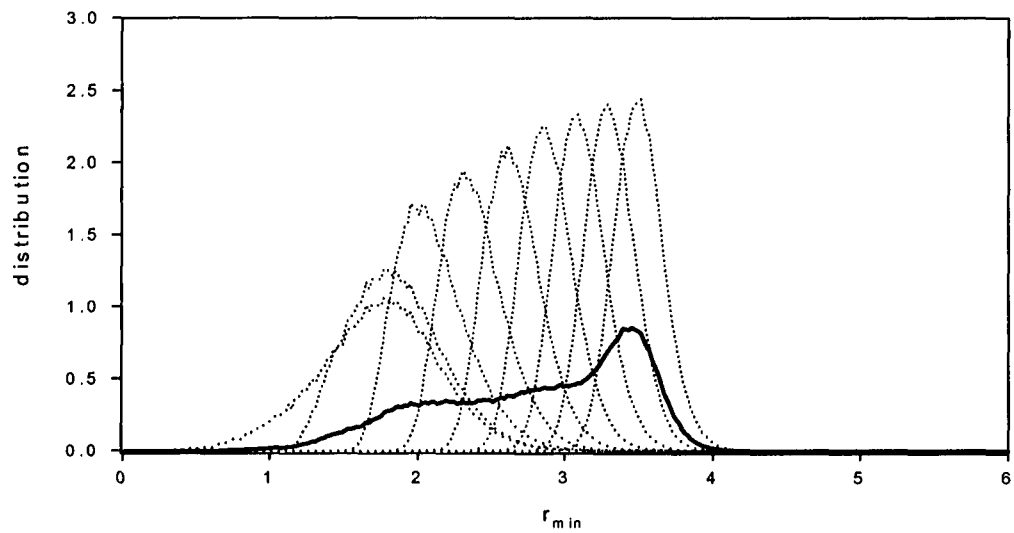
<그림 II-12> Distribution of the distance between the cavity and its nearest water molecule : results of the second simulation using p=2 for S0, S2, S4, S6 and S8 ($\Delta C_{m0}^{(2)} = \Delta F_{m0}^{(1)}$, $m = 2, 4, 6, 8$)



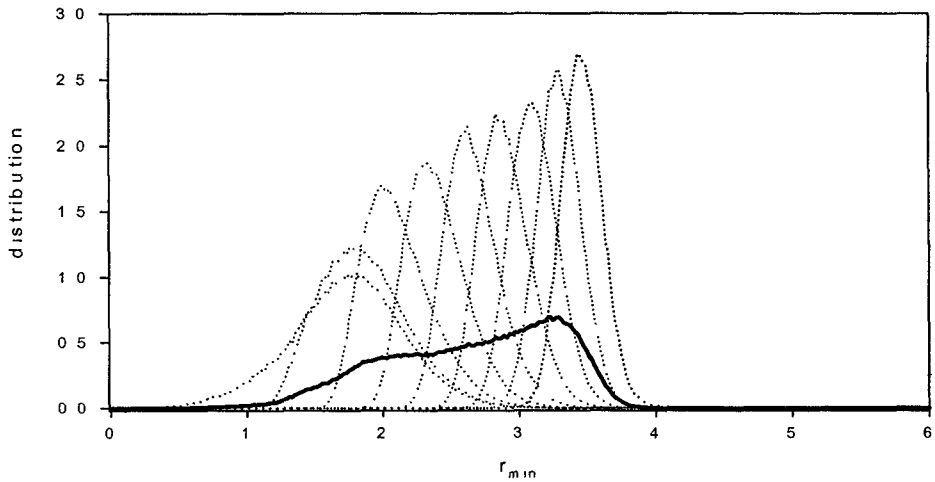
<그림 II-13> Distribution of the distance between the cavity and its nearest water molecule : results of the third simulation using p=2 for S0, S2, S4, S6 and S8 ($\Delta C_{m0}^{(3)} = \Delta F_{m0}^{(2)}$, $m = 2, 4, 6, 8$)



<그림 II-14> Distribution of the distance between the cavity and its nearest water molecule : results of the first simulation using p=2 for S0, S1, S2, S3, S4, S5, S6, S7 and S8 ($\Delta C_{m0}^{(1)} = 0,\ m = 1, 2, \cdots, 8$)



<그림 II-15> Distribution of the distance between the cavity and its nearest water molecule : results of the second simulation using p=2 for S0, S1, S2, S3, S4, S5, S6, S7 and S8 ($\Delta C_{m0}^{(2)} = \Delta F_{m0}^{(1)},\ m = 1, 2, \cdots, 8$)



<그림 II-16> Distribution of the distance between the cavity and its nearest water molecule : results of the third simulation using p=2 for S0, S1, S2, S3, S4, S5, S6, S7 and S8 ($\Delta C_{m0}^{(3)} = \Delta F_{m0}^{(2)}$, $m = 1, 2, \cdots, 8$)

또 창의 수가 너무 적으면, 각 예비시늉의 길이가 너무 길어지기 때문이다. 우리는 인접하는 두 창의 간격을 바꾸어가며 표본 추출된 배열들에 대한

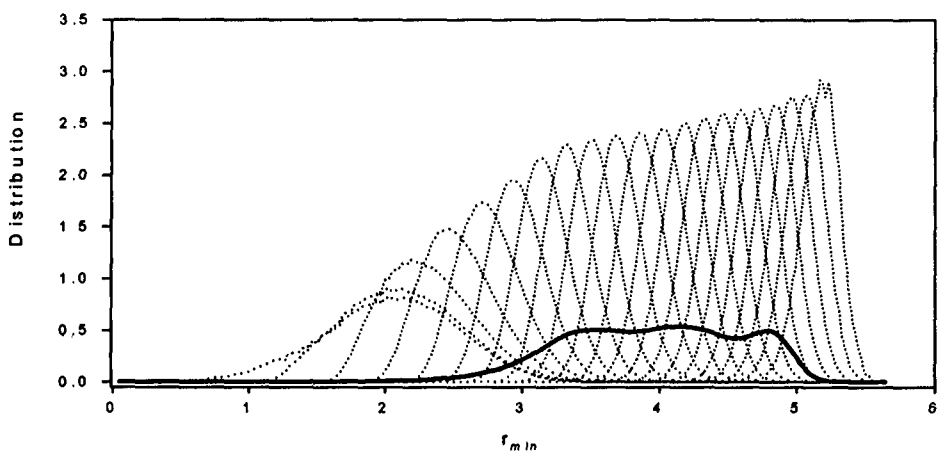
$$\ln\left[\frac{f_{m+1}}{f_m}\right] \qquad \text{(II-12)}$$

<table II-3> Free energies obtained from simulation with p = 0.5, 1, 2, and 4

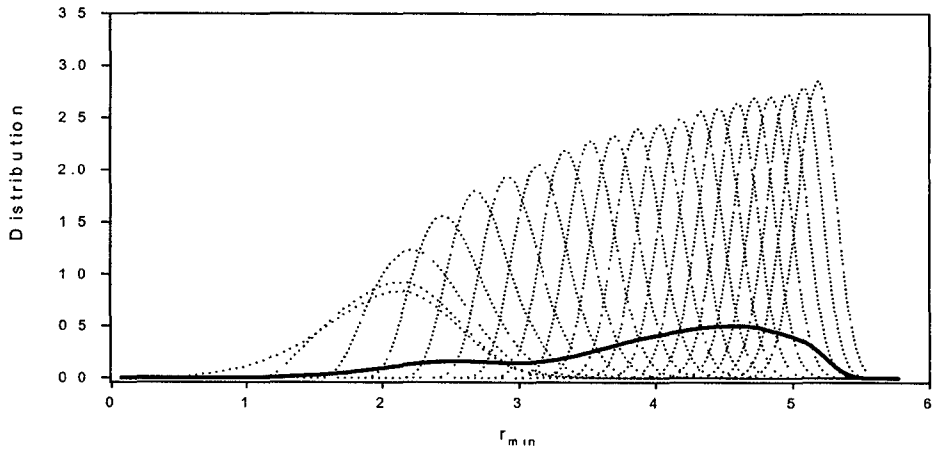
r_{th}	p = 05	p = 1	p = 2	p = 4
0	0.851	0.85	0.839	0.838
1.26	1.847	1.809	1.819	1.78
1.78	3.073	3.03	3.019	2.987
2.18	4.51	4.549	4.403	4.478
2.52	6.124	6.079	5.998	6.168
2.82	7.887	7.829	7.827	7.947
3.09	9.868	9.85	9.723	9.908
3.33	11.821	12.034	11.617	12.067
3.57	13.99	14.408	13.967	14.441
3.78	16.452	16.645	16.335	16.537

r_{th}	p = 05	p = 1	p = 2	p = 4
3.99	19.023	19.23	18.924	19.023
4.18	21.915	22.025	21.661	21.736
4.37	24.927	25.035	24.568	24.825
4.54	28.193	28.189	27.854	27.997
4.72	31.635	31.724	31.106	31.307
4.88	35.416	35.259	34.665	35.122
5.04	39.256	38.936	38.774	39.109
5.20	43.566	43.157	42.973	43.424
5.35	48.029	47.791	47.417	47.868

값의 분포를 관찰하였다. 한 순간배열에 대한 f_m의 값은 그 배열이 계 m의 배열공간영역에 속해 있는 정도를 나타내는 매개변수로 볼 수 있기 때문이다. 수식 II-6으로부터 알 수 있듯이, 추출된 순간 배열이 계 m에 전혀 속하지 않는다면 그 배열에 대한 f_m의 값이 0이 되는 반면, 계 m에만 속해 있다면 그 값은 1이 된다. 우리는 이웃하는 창들의 배열공간영역들이 서로 충분히 겹치도록 창들 사이의 간격을 정해야 한다. 배열 공간이 충분히 겹치는 경우에는 수식 II-12의 값은 0 근처에 주로 분포할 것이다. 계산결과, 창사이의 간격이 멀어질수록 수식 II-12의 값이 급격히 증가하였다. 이웃하는 창들 사이의 거리를 멀리 잡으면 예비 시늉에 대부분의 전산 시간을 허비하게 된다.



<그림 II-17> Distribution of the distance between the cavity and its nearest water molecule : results of the first simulation using p=1 for S0, S1,···, and S20 ($\Delta C_{m0}^{(3)} = \Delta F_{m0}^{(2)}$, $m = 1, 2, \cdots, 20$)



<그림 II-18> Distribution of the distance between the cavity and its nearest water molecule : results of the second simulation using p=1 for S0, S1,⋯, and S20 ($\Delta C_{m0}^{(3)} = \Delta F_{m0}^{(2)}$, $m = 1, 2, \cdots, 20$)

우리는 이웃하는 두 계들로 이루어진 쌍들에 대해 각각 10^6에서 4×10^6 까지 4회 예비시늉을 하여 자유에너지 값들을 계산하였다. p = 0.5, 1, 2, 4 사용한 그 최종적인 값들은 <table II-3>에 나열된 것과 같다. 이 값들을 사용하여 S0에서 S13까지의 계들을 한 번에 탐사하여 보았다. 각 p 값에 대해 10^8 배열들을 표본 추출하는 시늉을 두 번씩 하였다. 또한 p = 1과 2를 사용하여 <table II-1>에 나열된 모든 계를 탐사하여 보았다. 각 p 값에 대해 5×10^8 배열들을 표본 추출하는 시늉을 두 번씩 하였다. 이 실험들에서는 p값에 대한 의존도는 거의 없었다. <그림 II-17> 과 <그림 II-18>은 p=1을 사용하여 얻은 배열분포를 그린 것이다. 이 그림들에서 여실히 보여 지는 MES의 배열공간탐사능력에는 놀라울 뿐이다.

III. 결 론

본 연구에서 우리는 MES의 무게함수를 수식 II-7과 같이 일반화하여 '물 속에서의 공동 형성'의 여러 가지 경우에 적용하여 그 실용적 속성을 조사하였다. 두 개의 계만을 탐사한 경우에 대한 결과는 <그림 II-4> - <그림 II-10>을 통해 알 수 있듯이, 각 계가 차지하는 배열공간

상의 영역들이 서로 멀수록 그 두 영역을 공평하게 탐사하기가 어려워진다. 왜냐하면, 한 쪽 영역에서 다른 쪽 영역으로의 이동이 빈번하게 일어나지 못하게 되기 때문이다. 또한, p 값을 바꾸어 가며 한 계산 결과는, 작은 p 값을 사용하면 그 어려움이 더욱 심해지는 경향이 있음을 보여주고 있다. 그러므로 두 계가 너무 멀리 떨어져있는 경우에는 그 사이에 창들을 열어서 교량 역할을 해주는 중간 계들을 두어야한다. 이 경우에 대한 계산 결과들로부터 다음에 기술하는 것과 같은 MES의 속성을 알 수 있다. 두 계의 떨어짐 정도가 그렇게 심하지 않는 경우에는 모든 창들에 대한 시늉 매개변수인 ΔC_{ml}의 값들을 한꺼번에 조정해서 할 수 있으나, 심한 경우에는 불가능해지기 때문에 이웃 창들의 쌍에 대해 개별적으로 조정해야 한다. 열어야 할 창들이 수는 적당한 길이의 예비시늉으로 매개변수 ΔC_{ml}가 수렴된 값으로 조정이 될 정도이어야 한다. 그것은 수식 II-12 값의 분포로부터 결정할 수 있다. 0 근처에서의 분포 값이 봉오리 값의 절반 정도가 되도록 이웃하는 창들 사이의 간격을 잡으면 충분하다. 이렇게 창의 수가 결정되면, 다음과 같이 첫 번째 이웃 쌍에 대한 것부터 시작하여 매개변수 ΔC_{ml} 값을 차례차례 조정한다. 처음에 $\Delta C_{10} = 0$ 로 놓고 시작하여 계산되지는 ΔF_{10} 값으로 대치하는 것을 2-3회 반복한다. 그 다음 쌍에 대한 매개변수의 초기값을 바로 전 쌍의 시늉에서 계산된 자유에너지 값으로 잡아서 (즉, $\Delta C_{l,l+1} = \Delta C_{l-1,l}$) 조정을 반복한다. 이러한 체계로 $r_{th} \leq 5.64$ 의 공동들을 한꺼번에 탐사하는데 성공하였다. <그림 II-4> 에서 <그림 II-18>까지의 그림들은 놀랄만한 MES의 배열공간탐사능력을 증명해주고 있다. 본 연구의 결과들은 'MES를 다양한 생물분자 관련 분야에 적용하여 성과를 거둘 수 있다'는 믿음을 갖게 한다. 우리는 배열 공간에 대한 효율적 탐사가 절대적으로 요구되어지는 conformational changes 등은 물론이고, 단백질 접힘(protein folding) 연구에 활용하려 한다. 또한, MES 이용의 보편화를 위하여, 본 연구의 결과를 토대로 하여 MES를 패키지화할 계획이다.

참 고 문 헌

[1] T. L. Blundell, B. L. Sibanda, M. J. E. Sternberg, and J. M. Thornton, Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 326, 347 (1987)

[2] M. J. Sutcliffe, I. Haneef, D. Carney, and T. L. Blundell, Knowledge-based modeling of

homologous proteins, part I: Three dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* 1, 377 (1987)

[3] Harold A. Scheraga, Jaroslaw Pollardy, Adam Liwo, Jooyoung Lee, Cezary Czaplewski, Daniel R. Ripoll, William J. Wedemeyer, Yelena A. Arnautova, J.Comput.Chem. **23**, 28-34(2002).

[4] Adam Liwo, Piotr Arlukowicz, Cezary Czary Czaplewski, Stanislaw Oldziej, Jaroslaw Pillardy, and Harold A. Scheraga, PNAS **99**, 1937-1942(2002).

[5] Jorgensen, W.; Ravimohan, C. J Chem Phys 1985, 83, 3050.

[6] Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A. Science 1987, 235, 574.

[7] Jorgensen, W. L.; Nguyen, T, B, J. Comput. Chem. 1993, 14, 195-205.

[8] Singh, U. C.; Weiner, P. K.; Caldwell, J.; Kollman, P. *Amber 3.0*, University of California: San Francisco, 1986

[9] Pearlman, D.; Case, D.; Caldwell, J; Singh, U. C.; Weiner, P. K.; Kollman, P. *Amber 4.0*, University of California: San Francisco, 1992

[10] Jorgensen, W. L. *BOSS*, version 3.2, Yale University: New Haven, 1992

[11] Kollman, P. *Chem Rev* **1993,** 93, 2395.

[13] Reynolds, C. A.; King, P. M.; Richards, W. G. *Mol. Phys.*, **1992,** 76, 251.

[14] Singh, U. C.; Brown, F. K.; Bash, P. A.; Kollman, P. A. *J. Am. Chem. Soc.*, **1987,** 109, 1607.

[15] Valleau, J. P.; Card, J. *J Chem Phys* **1972**, 57, 5457.

[16] Torrie, G.; Valleau, J. P. *Chem Phys Lett* **1974**, 28, 578.

[17] Torrie, G.; Valleau, J. P. *J Comp Phys* **1977**, 23, 187.

[18] Valleau, J. P. *J Chem Phys* **1993**, 99, 4718.

[19] Han, K.-K. *Phys. Lett. A*, **1992**, 165, 28.

[20] Han, K.-K. *Phys. Rev. E*, **1996**, 54, 6906.

[21] Han, K.-K.; Kim, K. H.; Mhin, B. J.; Son, H. S. *J. Compu. Chem.* **2001**, 22, 1004.

[22] Berg, B. A.; Neuhaus, T. Phys Lett 1991, B267, 249-253.

[23] Berg, B. A.; T. Phys Rev Lett 1992, 68, 9-12.

[24] Lee J. Phys Rev Lett 1993, 71, 211-214.

[25] Lee, J. Phys Rev Lett 1993, 71, 2353.

[26] Lyubartsev, A. P.; Martinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. J Chem Phys 1992, 96, 1776-1783.

[27] Marinari E.; Parisi, G. Europhy Lett 1992, 19, 451-258.

[28] Hukushima, K.; Nemoto, K. J Phys Soc Jpn 1996, 65, 1604-1608.

[29] Hukushima, K.; Takayama, H.; Nemoto, K. Int J Mod Phys C 1996, 7, 337-344

[30] Geyer, C. J. In Computing Science and Statics; Proceedings of the 23rd Symposium on the Interface; Keramidas, E. M., Ed; Interface Foumdation; Fairfax Station, 1991, pp 156-163