

Developing Virtual Screening Program for Lead Identification

선도화합물 탐색을 위한 고효율가상탐색 프로그램 개발

Ky-Youb Nam¹, Yongkee Cho¹, Chang-Joon Lee², Jae Hong Shin¹, Jung Won Choi³, Joon-Min Gil³, Hark Soo Park³, Il Sun Hwang³, Kyoung Tai No^{2*}

¹ Drug Discovery Division, Research Institute of Bioinformatics and Molecular Design Seoul, Korea

² Department of BioTechnology, Yonsei University, Seoul, Korea

³ Supercomputing Center, Korea Institute of Science and Technology Information

*To whom correspondence should be addressed. E-mail: ktno@bmdrc.org

Abstract

The docking and *in silico* ligand screening procedures can select small sets of lead -like candidates from large libraries of either commercially or synthetically available compounds; however, the vast number of such molecules make the potential size of this task enormous. To accelerate the discovery of drugs to inhibit several targets, we have exploited massively distributed computing to screen compound libraries virtually. The Korea@HOME project was launched in Feb. 2002, and one year later, more than 1200 PC's have been recruited. This has created a 31 -gigaflop machine that has already provided more than 1400 hours of CPU time. It has allowed databases of millions of compounds to be screened against protein targets in a matter of days. Now, the virtual screening software suitable for distributed environments is developed by BMD. It has been evaluated in terms of the accuracy of the scoring function and the search algorithm for the correct binding mode.

Introduction

인간 유전체 프로젝트(Human Genome Project)를 통해 많은 양의 유전체 정보가 얻어짐에 따라 각종 질병에 대한 다양한 표적 단백질(target protein)이 제시되고 있다. 유전자 지도의 완성에 따라 기존에 400 개 정도

에 해당하던 신약개발의 표적 단백질이 약 3 만개 정도로 급격히 증가할 것으로 예상하고 있으며, 이 말은 다시 말하면 질병을 치료하는 방법이 다양해지며, 인간 개개의 질병 원인에 따른 맞춤형신약에 대하여 연구가 가능해질 것이라 예기할 수 있다.

기존에 사용하던 분자 도킹(Molecular Docking)방법을 통해 한 표적 단백질에 대해 가상 탐색(Virtual Screening)을 실시할 때

This work is supported by Ministry of Information and Communication, and Korea@Home project.

현재 컴퓨터 계산 능력을 고려하면 많은 시간이 걸리며 결과를 분석하는 작업 또한 만만치 않았다. 더구나 인간 계층 지도 분석을 통해 늘어난 많은 질병 관련 표적 단백질에 대해서 모두 가상 탐색을 실시하려면 현재 계산 능력이 도저히 감당할 수 없는 정도의 계산 양이 산출 될 수 있다. 이러한 계산 능력의 한계를 극복하기 위해 수 년 전부터 분산 컴퓨팅이 대안으로 제시되고 있으며 상업적으로 이미 사용이 가능해져 있는 실정이다.

2003 년 발표에 따르면 우리나라는 초고속 인터넷 보급률 1 위와 1 인당 인터넷 사용시간도 1 위를 달리고 있다. 하지만 이러한 우수자원에도 불구하고 대부분의 인터넷 사용자가 정보 검색, 게임, 메일 확인 같은 소극적인 이용에 머물러 있는 실정이고, 자원의 90% 이상이 유휴자원으로 낭비되고 있다고 한다. 이처럼 우수한 인터넷 환경과 유휴 PC 자원을 잘 활용하면, 짧은 시간 동안 대규모 데이터를 처리해야 하는 경우 해결책을 제시해 줄 수 있다.

국내에서는 이러한 한계를 극복하기 위해 Korea@HOME 이란 분산처리 컴퓨팅 환경 구현 및 응용에 대한 프로젝트가 시작되었으며, 이로 인해 인터넷에 접속되어 있는 유휴 PC 자원을 이용하여 이러한 한계를 극복할 수 있게 되었다. 외국에서는 이미 이와 비슷한 다양한 프로젝트가 시작되었는데 외계에 생명체를 찾아 내고자 하는 SETI@HOME 프로젝트[1]나 단백질의 3 차원 접힘 문제와 관련된 Folding@HOME 프로젝트[2]나, AIDS 로 잘 알려진 후천성 면역결

핍증의 치료제를 찾아내는 FightAIDS @HOME 프로젝트[3] 등이 있다.

국내에서는 암관련 표적 단백질로 잘 알려져 있는 CDK2(Cyclin-Dependent Kinase 2) 를 대상으로 하여 신약 후보 물질을 찾는 응용과제가 처음으로 시작되었다. 이 논문은 Korea@HOME 네트워크를 이용한 가상 탐색과 그 알고리즘에 대한 설명, 그리고 프로그램의 검증 과정을 기술할 것이다. 이 검증과정은 기존에 이미 잘 알려진 단백질-리간드 x-ray crystal 결합 구조를 예측하여 비교함으로써 프로그램의 정확도를 측정하는 것을 포함한다. 여기에 사용한 단백질-리간드 결합구조는 Dihydrofolate Reductase (DHFR), Streptavidin, CDK2 x-ray 결합구조를 사용하였으며, 본 연구실에서 개발된 단백질-리간드 토크프로그램으로 예측능력을 살펴보았다.

Methods

경험적 결합세기 함수

Hydration Free Energy Density Model

단백질-리간드 도킹의 연구를 수행함에 있어서 단백질과 리간드가 실제로 존재하는 시스템인 용매에 대한 효과를 고려하는 것은 매우 중요하다. 용액에 대한 연구를 효과적으로 묘사하는 수화 자유에너지 밀도 (Hydration Free Energy Density)[4] 모델을 적용했다. 수화 자유에너지 밀도 모델에서 용매와 용질의 상호 작용은 각 부분 상호 작용의 합으로 나타낸다. 이 때 각 부분 상호 작용은 물리적으로 의미 있는 매개변수를 사용하며, 이들은 원자의 전하량(partial

charge)과 편극도(polarizability)로부터 얻어질 수 있다.

비결합 상호작용(Nobonding Interaction)

단백질 분자와 리간드 분자간의 상호작용은 비결합 에너지(Nonbonding energy)로 나타내어 질 수 있다. 비결합 에너지는 van der Waals 에너지와 정전기적(electrostatic) 에너지 등으로 나타내어 질 수 있다. 여기서는 cvff 힘장(ForceField)을 사용하였고,[5] 비결합 상호 작용은 van der Waals 에너지와 정전기적 인력의 합으로 나타내었다. 이때 거리 11.5Å 이상 일 때는 스위칭 함수를 적용하였다.[6]

경험적 결합 세기 함수

경험적 결합세기 함수는 아래와 같이 정의하였다.

$$\Delta G_{bind} = \Delta G_{vdw} + \Delta G_{elec} + \Delta G_{HFED} + \Delta G_{rot} + \Delta G_{M,W} \quad (1)$$

위의 (1)식에서, van der Waals 에너지와 정전기적 에너지는 기본적인 분자역학 term 으로서, 각각 Lennard-Jones 12-6 dispersion/repulsion term 과 Coulombic electrostatic potential term 을 이용하였다. ΔG_{rot} 는 리간드의 회전이 가능한 이면각의 총 개수, $\Delta G_{M,W}$ 는 리간드의 분자량과의 상관관계를 나타낸다. ΔG_{HFED} 는 자유 수화에너지 총합을 나타내는 항이다.

기존에 생리활성 값이 밝혀진 130 개의 단백질-리간드 결합구조 중에서 동종의 단백질을 제외한 52 개의 대표적인 단백질을 선

택했다. 위의 식을 통해 계산된 결합에너지 값과 생리활성과의 상관관계를 조정(Fitting)한 결과 그림 1 과 같이 R^2 값이 0.6008 까지 되도록 VA10A 최적화 알고리즘을 사용하여, 위 (1)식의 계수를 결정했다.[7-12]

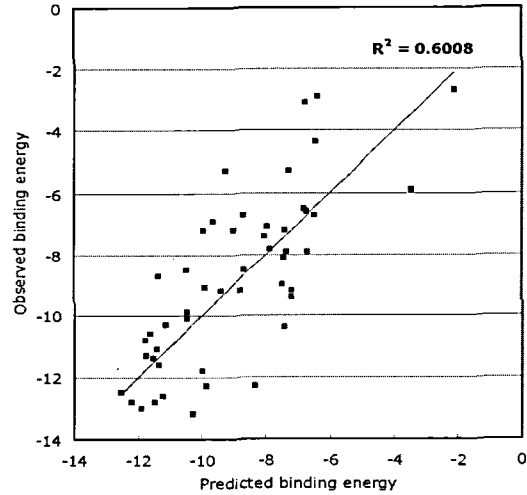


그림 1) 경험적 결합에너지 함수 조정, 실험 값과 계산값의 비교.

단백질 활성자리 계산

단백질의 x 선 구조를 이용하여 리간드가 결합할 수 있는 자리를 찾는 방법이다.[13] 이 방법은 빠른시간내 단백질의 빈공간을 탐색하는 유용한 방법이다.

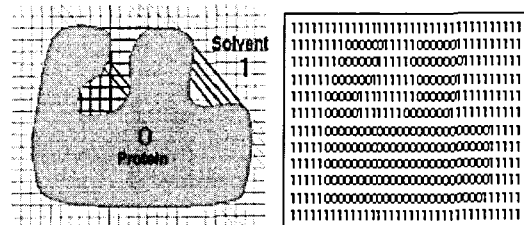


그림 2) 단백질의 격자표현과 각 격자점 값 표현

Real-valued representation

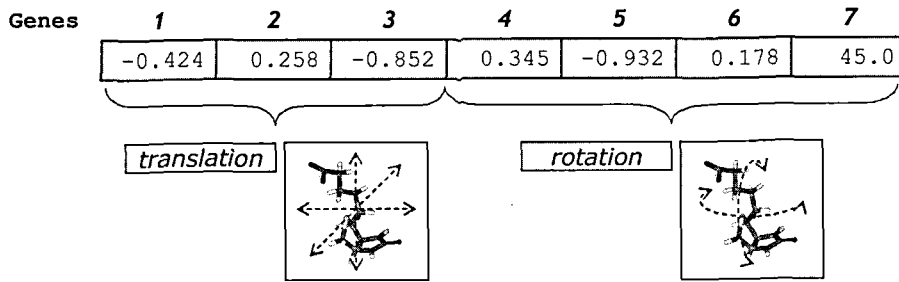


그림 4) 실제값으로 인코딩된 유전자. 1gene 부터 3gene 은 분자의 이동(translation)을 나타내며 4gene 부터 7gene 은 분자의 회전(rotation)을 나타냄.

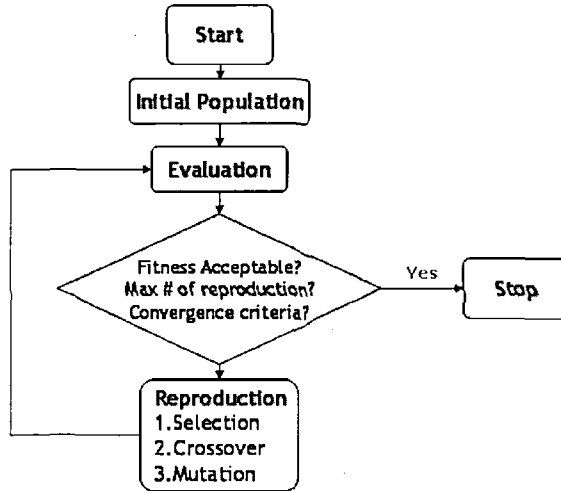


그림 5) 진화 알고리즘의 흐름도

즘에서 Fitness 값은 각 응용에 따라 그 계산 방법이 다른데, 일반적으로 이는 현재 상태를 나타내는 상대적인 양이 될 것이다. 본 응용에서는 cvff 힘장을 기반으로 한 경험적 결합세기 값을 사용하였는데 이 값은 각 리간드에 대한 실제 생리 활성의 값과 맞도록 조정되었다. 이 에너지 값을 아래와 같은 (2)식으로 다시 조정하여 Fitness 값으로 사용하였으며, 아래의 (2)식을 이용하면 알고리즘이 진화하는 동안 Fitness 값은 0에서 1로 접근 하도록 각 유전자를 결정하게 된다.

$$\rho = \begin{cases} \frac{|S_{\max} - S_{\text{curr}}|}{|S_{\max} - S_{\min}|} & \text{if } S_{\text{curr}} > S_{\min} \\ 1 & \text{if } S_{\text{curr}} \leq S_{\min} \end{cases} \quad (2)$$

위 (2)식에서, S_{\max} , S_{\min} , S_{curr} 은 각각 결합에너지의 최대값, 최소값, 현재값을 나타낸다.

$$F = \rho^{S_{\text{ave}}} \quad (3)$$

여기서 F 는 Fitness값을 S_{ave} 는 결합에너지의 평균값을 나타낸다.

화합물 데이터 준비

리간드 화합물 데이터베이스

본 과제에 사용된 화합물은 약 500,000 개의 화합물이 사용되었다. 초기 화합물은 상업적으로 유용한 화합물 데이터베이스로부터 시작했다.(Asinex, Bionet, Chemdiv, CStar, MDDR, Maybridge, Tripos) 이 화합물들을 다 합치면 약 100 만개가 넘었다. 이 화합물들을 분자량이나 특이 원소 포함 여부에 따라 필터링하여 약 60 만개의 화합물을 얻을 수 있었다. 보통 화합물 데이터 베이스는 2D 구조 데이터베이스인데 가상 탐색을 하려면 3D 구조가 있어야 하므로 Corina 를 사용해서 3D 로 변환하였다.[16] 안정한 구조를 계산하기 위해 Sybyl 을 사용하여 최적화 시켰다. Chemoinnovation Chem4D 의 Nomenclator database 는 화합물 데이터를 IUPAC name 으로 변환해 주는 기능을 가졌는데 이를 이용하여 화합물의 이름을 첨가하였다.[17] 각각 화합물들이 가지고 있는 여러 가지 물리화학적 성질이나 표현자들을 계산하기 위해 PreADME 를 사용하였다.[18] ADME 란 Absorption, Distribution, Metabolism, Excretion 등을 의미하는 말로서 각 화합물에 대한 흡수, 분포, 대사, 배설 등에 대한 여러 가지 지표들이 사용되어진다.

단백질 데이터

그 다음에는 크로스오버(Crossover)와 변이(Mutation) 과정을 거치게 되어 최적화된 리간드의 구조로 최적화 되어진다.

타겟 단백질의 구조는 Protein Data Bank(단백질 구조 은행)[19]에서 가져왔으며 InsightII 에서 수소를 첨가하였고, 실제 생체 내의 상태와 일치시키기 위해 pH 를 7.4 로 조정하였다. 수소의 첨가로 인한 반발력을 최소화하기 위하여 InsightII/Discover 프로그램을 사용하여 구조 최적화를 수행하였다.

화합물 정보 데이터

화합물의 데이터는 preADME 를 통해 계산되며 계산된 값은 MDL connection table 형태로 저장된다.[20] MDL MOL 파일은 사용자 지정 데이터 필드의 삽입이 용이할 뿐만 아니라 다중 화합물 정보도 포함 할 수 있기 때문에 널리 사용되고 있다. 하지만 순차적으로 필드를 검사해야 하며 프로그램마다 필드 값을 기술 하는 방법이 조금씩 차이가 나기 때문에 서로 다른 화합물 데이터를 혼합하는데 문제가 있을 수 있다. 하지만 데이터 기술의 표준으로 자리 잡고 있는 XML 을 도입하면 호환성과 객체의 직렬화 면에서 유리해질 것으로 전망한다. 현재 응용프로그램은 데이터 게시 목적으로만 XML 을 사용하고 있지만, CML 이나 XSD 같은 화합물 구조의 표준 XML 파일 형식 또는 그 확장을 도입하여야 할 것이다.[21]

Results

본 논문에서 제안한 진화 알고리즘을 이용한 가상 탐색 기술을 검증하기 위하여 다

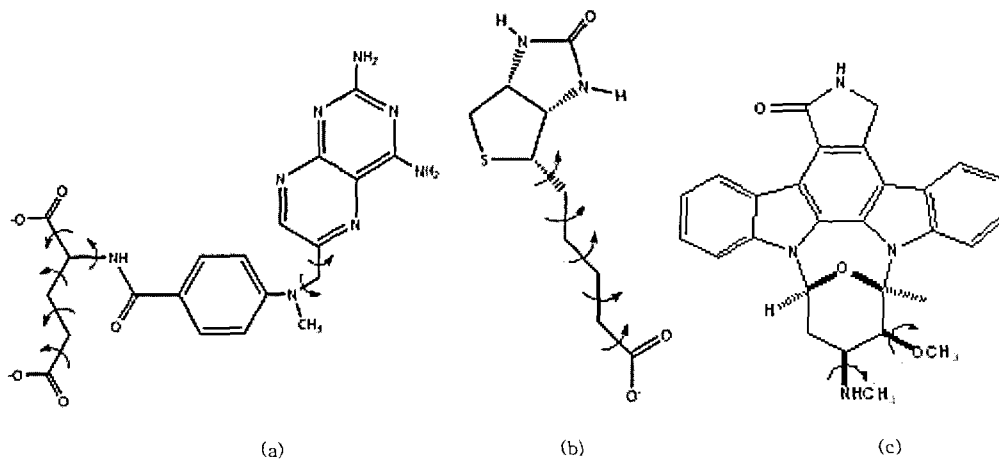


그림 6 분자 도킹에 사용된 리간드 : (a) methotrexate, (b) biotin 과 (c) staurosporin

<표 1> 분자 도킹으로 얻어진 최소에너지의 구조와 실험값과의 비교

Energy(kcal·mol ⁻¹) and RMSD Å					
PDB code	RMSD	ΔG_{vdw}	ΔG_{elec}	ΔG_{HFED}	ΔG_{pred}
4DFR	0.57	145.88	-1.01	-171.87	-6.66
1STP	1.14	-23.68	-8.19	-43.02	-7.02
1QMZ	1.16	470.29	-6.59	-92.36	44.81

음과 같은 실험을 하였다. 잘 알려진 단백질-리간드 결합구조에 대해 분자도킹에 의한 구조와 실제 결합구조간의 비교를 통해 프로그램을 검증하였다. 분자 도킹실험에 사용된 단백질-리간드는 Dihydroforate reductase-methotrexate (4DFR), Streptavidin-biotin(1STP)과 CDK2-staurosporin(1QMZ)가 사용되었다.

위 <표 1>은 계산값과 실제 생리활성의 차이, 그리고 RMSD(Root mean square deviation)을 나타낸다. 위의 <표 1>에서 볼 수 있듯이 RMSD 값의 차이가 1.5Å 미만

의 값을 보였다. DHFR-methotrexate (4DFR)의 경우 0.57 Å, Streptavidin-biotin(1STP)의 경우 1.14 Å, 그리고, CDK2/staurosporin(1QMZ)의 경우 1.16 Å 의 RMSD 를 보여주고 있다. 이는 거의 X-ray 결정 구조와 분자 도킹의 최소 에너지 구조와 거의 일치함을 보여주고 있다. CDK2/stausporine 의 경우 일반적으로 stausporine 이 induced fit 을 일으키는 대표적인 inhibitor 로 단백질의 구조를 많이 변화시키는 것으로 잘 알려져 있으며, 본 실험에서 많은 van der Waals repulsion 의 크기가 470.29 kcal/mol 로 매우 크게 작용하고 있음을 살펴 볼 수 있다.

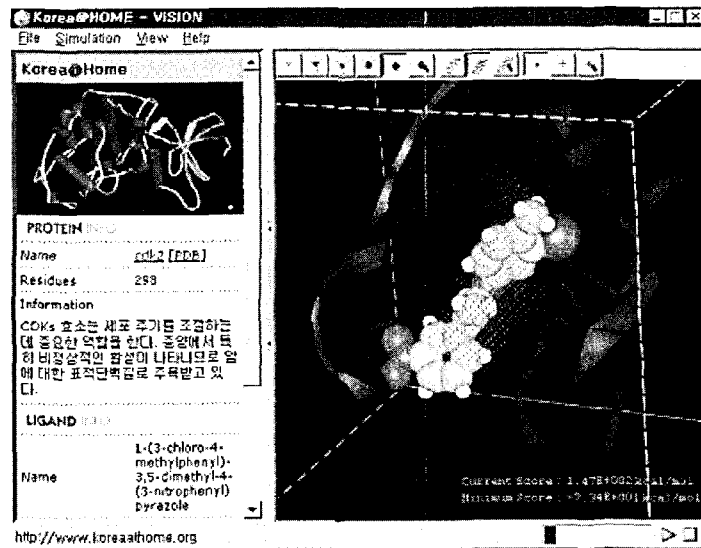


그림 7. 가상 탐색 프로그램 GUI - VISION

이러한 문제는 단백질-리간드의 hardcore potential 을 사용함으로써 발생하는 문제이며, 매우 작은 공간에 큰 리간드가 결합함으로써 생기는 van der Waals repulsion 의 크기를 줄이는 방법을 향후에 테스트 되어져야 할 것이다.

본 단백질-리간드 예측방법에서는 HFED 방법을 사용하여, 단백질-리간드 결합에서 발생하는 수화에너지에 대한 고려를 수행하였으며, 이러한 수화에너지에 대한 고려 방법은 Poission-Boltmann equation, Generalized Born model 등과 같은 implicit water 모델이 많은 docking 프로그램에서 사용 중에 있다.

Discussion

우리는 가상 탐색 프로그램인 VISION 의 개발을 위하여, 단백질-리간드 52 개의 대표적인 단백질을 선택하여, 계산된 결합에너지 값과 생리활성과의 상관관계를 조정

(Fitting)한 결과 R^2 값이 0.6008 까지 되도록 VA10A 최적화 알고리즘을 사용하여 경험적 결합세기 함수를 설계하였고, 단백질의 활성부위를 효과적으로 찾을 수 있는 Manfred 알고리즘을 재현하여 단백질 활성부위의 모양을 결정하는 방법을 구축하였으며, 리간드의 회전에 관한 4 개의 변수를 갖는 쿼터니온(Quaternion) 방법과 리간드의 움직임에 관한 3 개의 변수를 gene 들로 사용하는 유전자알고리즘을 사용하여, 단백질 내에 리간드의 최적화된 구조를 예측하였다. 이러한 유전자 알고리즘은 Monte carlo 방법보다 일반적으로 시간과 global optimum 을 찾는 데 유리하다고 알려져 있으며, 최종적으로 검증과정을 통해서 4DFR 의 경우에는 0.57Å, ISTP 의 경우에는 1.14Å, 그리고 1QMZ 의 경우에는 1.16Å 의 RMSD 의 정확도를 볼 수 있었다. 향후 더 많은 종류의 다양한 단백질-리간드 결합구조에 대한 검증을 수행할 예정이다. 또한 더욱 다양한 단백질-리간드 결합구조를 protein data bank 에서 추

출하여, 광범위하게 사용 가능한 결합세기 예측프로그램을 보완해야 하며, 많은 수의 리간드에 대한 고효율가상탐색을 수행하기 위해서는 단백질과 리간드의 1 대 1 계산 보다는 미리 단백질과 grid point 의 상호작용을 계산하여 보다 빠른 시간에 단백질-리간드 결합구조를 예측하는 알고리즘으로 개선되어야 할 것이다.

그림 7 은 VISION 프로그램의 GUI 를 보여주는 그림이며, 단순히 prompt 창에서만 실행이 되는 것이 아니라, 현재 진행되고 있는 프로젝트, 타겟 단백질, 리간드에 대한 정보를 보여줄 수 있도록 설계하였다. Korea@HOME 을 통해 얻은 신약 후보 물질은 좀더 정밀한 검증과정을 거쳐 False Positive 등을 걸러 내게 된다. 많은 양을 검색해야 하므로 빠르지만 다소 부정확한 방법을 사용하였기 때문에, 이 단계에서는 좀더 정확한 방법을 취해 활성 여부를 더 자세히 예측할 수 있게 된다. 이 단계를 거쳐 얻어진 물질들이 실제로 활성을 갖고 있는지는 생물학적 검증 과정을 거쳐야 한다. 이 과정을 통과해야 신약 후보 물질로 검증될 수 있다.

Pentium IV 한 대의 컴퓨터로 한 타겟 단백질에 대해서 50 만개의 화합물을 모두 가상적으로 탐색하려면 평균적으로 3 년 정도의 시간이 걸릴 것으로 예상되지만 비교적 규모가 큰 각 제약회사 마다, 인간 게놈 프로젝트로부터 얻어진 자기 자신들의 독자적인 타겟 단백질에 대한 신약 후보물질을 검색하는데 분산 컴퓨팅 환경을 이용하거나 도입할 것을 심각하게 고려하고 있는 실정이다. Korea@HOME 네트워크를 통해 분산

하여 처리하여 몇 달 또는 몇 일만에 계산을 끝낼 수 있게 되었다. 이러한 이유로 현재 Korea@HOME 이 더 규모가 커지고 더 많은 계산량을 수용할 수 있게 되면 유연한 (Flexible) 리간드에 대해 계산하는 것도 가능해지고 현재 보다 정밀한 계산 방법을 사용할 수 있게 될 것이다. 앞으로 신약개발을 위한 가상 탐색에 보다 다양한 알고리즘이 적용될 예정이며 새로운 표적 단백질에 대해서도 적용하려고 한다.

Acknowledgements

본 연구는 정보통신부, 한국과학기술정보연구원 korea@Home 사업에 의해 지원 되었음.

References

- [1] <http://setiathome.ssl.bekeley.edu>
- [2] <http://www.stanford.edu/group/pandegroup/folding>
- [3] <http://fightaidsathome.scripps.edu>
- [4] K. T. No *et al.*, *Biophys. Chem.*, **78**, 127-145(1999)
- [5] Hwang, *et al.*, *J. Amer. Chem. Soc.*, **116**, 2515-2525(1994)
- [6] ForceField-Based Simulation, MSI(1997)
- [7] Morris, G. M. *et al.*, *J. Comp. Chem.*, **19**, 1639-1662(1998)
- [8] Eldridge, M. D., *et al.*, *J. Comput. Aided Mol. Des.*, **11**, 425-445(1997)
- [9] Bohm, H. J., *J. Comput. Aided Mol. Des.*, 383-396(1998)
- [10] Murray, C. W., *et al.*, *J. Comput. Aided Mol. Des.*, **13**, 547-562(1999)

- [11] Oprea, T., *et al.*, *Persp. in Drug Discovery and Design*, **9/10/11**, 36-61(1998)
- [12] Wang, R., *et al.*, *J. Mol. Model.*, **4**, 379-384(1998)
- [13] Manfred H., *et al.*, *J. Mol. Graph. and Model.*, **15**, 359-363(1997)
- [14] David E., *Evolutionary Algorithms in Molecular Design*, Wiley-VCH(2000)
- [15] F. Dunn, *et al.*, *3D Math Primers for Graphics and Game Development*, Worldware(2002)
- [16] Gasteiger, *et al.*, *Tetrahedron Comp. Method*, **3**, 537-547(1990)
- [17] <http://www.cheminnovation.com/products/nomenclature.asp>
- [18] <http://preadme.bmdrc.org>
- [19] <http://www.rcsb.org>
- [20] <http://mdli.com/downloadspublic/ctfile/ctfile.jsp>
- [21] <http://www.xml-cml.org>