Bioinformatics and Biosystems 2006, Vol. 1, No. 1, pp. 82-87

단백질 서열정렬 정확도 예측을 위한 새로운 방법

(A new method to predict the protein sequence alignment quality)

이민호, 정찬석, 김동섭

한국과학기술원 바이오시스템학과

초 록

현재 가장 많이 사용되는 단백질 구조 예측 방법은 비교 모델링 (comparative modeling) 방법이다. 비교 모델링 방법에서의 정확도를 높이기 위해서는 alignment의 정확도 역시 매우 필수적으로 필요하다. 비교 모델링 과정 중의 fold-recognition 단계에서 alignment의 정확도에 의해 template을 고르는 방법은 단지 가장 비슷한 template을 선택하는 방법에 비해 주목을 받지 못하고 있다. 최근에는 두 가지의 alignment에 사이의 shift 정보를 바탕으로 한 shift score라는 수치가 alignment의 성능을 표현하기 위해서 개발되었다. 우리는 더 정확한 구조 예측의 첫걸음이 될 수 있는 shift score를 예측하는 방법을 개발하였다. Shift score를 예측하기 위해 support vector regression (SVR)이 사용되었다. 사전에 구축된 라이브러리 안의 길이가 n 인 template과 구조를 알고 싶은 query 단백질 사이의 alignment는 $n+2$ 차원의 input 벡터로 변환된다. Structural alignment가 가장 좋은 alignment로 가정되었고 SVR은 query 단백질과 template 단백질의 structural alignment과 profile-profile alignment 사이의 shift score를 예측하도록 training 되었다. 예측 정확도는 Pearson 상관계수로 측정되었다. Training 된 SVR은 실제의 shift score와 예측된 shift score 사이에 0.80의 Pearson 상관계수를 갖는 정도로 예측하였다.

Abstract

The most popular protein structure prediction method is comparative modeling. To guarantee accurate comparative modeling, the sequence alignment between a query protein and a template should be accurate. Although choosing the best template based on the protein sequence alignments is most critical to perform more accurate fold-recognition in comparative modeling, even more critical is the sequence alignment quality. Contrast to a lot of attention to developing a method for choosing the best template, prediction of alignment accuracy has not gained much interest. Here, we develop a method for prediction of the shift score, a recently proposed measure for alignment quality. We apply support vector regression (SVR) to predict shift score. The alignment between a query protein and a template protein of length n in our own library is transformed into an input vector of length $n+2$. Structural alignments are assumed to be the best alignment, and SVR is trained to predict the shift score between structural alignment and profile-profile alignment of a query protein to a template protein. The performance is assessed by Pearson correlation coefficient. The trained SVR predicts shift score with the correlation between observed and predicted shift score of 0.80.

서 론

초기의 CASP (Critical Assessment of Structure Prediction) 실험에서는 ab initio method들이 주를 이루었지만 knowledge-based 방법, 즉, 단백질 비교 분석 모델링 (comparative protein structure modeling) 방법은 최근의 실험에서 훌륭한 결과를 보이고 있다. (Kinch et al., 2003; Moult, 2005). 일반적으로 비교 모델링 방법은 구조를 알지 못하는 단백질과 가장 가까운 fold를 찾아내는 단계인 fold recognition이라는 과정을 포함하고 있다. 그 과정이 끝난 후에는 선택된 fold와 target 단백질 사이의 가능한 alignment가 만들어진다. 마지막으로 선택된 fold template과 가능한 alignment 중 가장 좋은 것들이 MODELLER (Marti-Renom et al., 2000; Sali and Blundell, 1993)와 같은 모델링 프로그램에 의해 target 단백질의 구조를 만드는데 사용되게 된다. Homology 모델링 프로그램들이 다른 정보가 아닌 오직 alignment만을 사용한다는 점에서 target과 template사이의 양질의 alignment는 성공적인 homology 모델링에 있어서 필수적이다.

Fold recognition 방법은 일반적으로 두 가지 종류로 분류될 수 있다. 하나는 Hidden Markov model (HMM) 방법 (Karplus et al., 1999) 과 PSI-BLAST (Altschul et al., 1997)

교신저자: 김동섭 (Email: kds@kaist.ac.kr)

This work is supported by grant number M1052900000205-N290000210 from Ministry of Science and Technology of Korea. Computational resources are provided in part by the IBM SUR Grant.

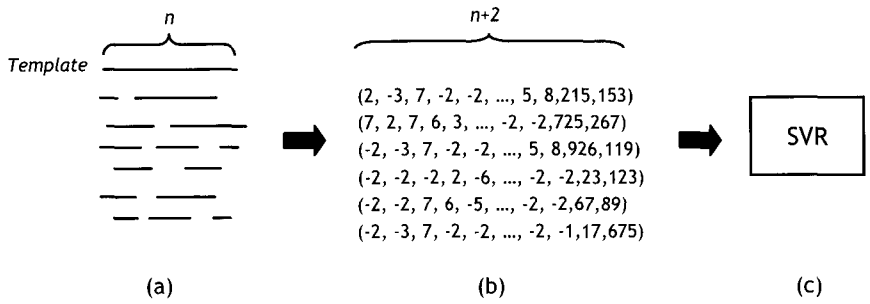

그림 1. Alignment로부터 input feature vector를 만드는 과정 (a) 길이가 n인 template 서열이 profile-profile alignment 방법으로 다른 sequence example들에 정렬된다. (b) 각각의 alignment는 n개의 위치에서의 alignment 스코어와 전체 스코어, query 서열의 길이로 이루어지는 (n+2) 차원의 feature vector로 변환된다. (c) 이 feature vector들이 각각의 template에 대한 SVR을 training 하는 데에 사용된다. Han et al. (2005)의 그림으로부터 수정되었다.

과 같이 오직 서열 정보만을 사용한다. 반면에 GenTHREADER (Jones, 1999) 나 PROSPECT (Kim et al., 2003; Xu and Xu, 2000) 와 같은 방법들은 서열 정보뿐만 아니라 구조 정보를 함께 사용한다.

하지만 어떠한 방법이 사용된다 할지라도 현재의 fold recognition 방법은 단지 target 단백질과 가장 비슷한 template를 어떻게 찾아내야 할지에 초점을 맞추고 있다. Jones (1997)는 이미 비교 모델링에서 alignment의 정확도의 중요성을 지적하고 있었다. 사실 homology 모델링에서 가장 중요한 요소는 target과 template의 유사도가 아니라 alignment의 정확도이다. 예를 들어 target 단백질과 가장 유사한 template가 fold recognition 단계에서 선택되었다고 하더라도 실제의 alignment 정확도는 다른 template과의 alignment보다 나쁠 수 있다. 그러나 높은 정확도를 갖는 alignment를 얻기 위해서는 target과 template 모두의 구조적 정보가 사용 가능하여야 한다. 이러한 사실은 더 정확한 fold recognition을 위해 어떠한 구조적 정보 없이 alignment의 정확도를 예측하는 새로운 방법의 필요성을 암시한다.

Alignment의 중요성 때문에 alignment reliability를 유추하거나 가장 좋은 alignment를 찾는 많은 접근방법이 개발되어 왔다. 이러한 방법에는 Holmes and Durbin (1998)가 고안한 방법과 sum-of-pairs score (Thompson et al., 1999) , 그리고 Sauder, et al. (2000)에 의해 modeler's score 라고도 불리는 reverse sum-of-pairs score(Edgar and Sjolander, 2004) 등이 있다. 최근에는 shift score라고 이름 지어진 새로운 척도가 기존에 개발된 스코어들의 단점을 해결하기 위해서 제안되었다 (Cline et al., 2002). Sum-of-pairs score는 over-alignment에 대하여 penalize하지 못하고, reverse sum-of-pairs score는 under-alignment에 대하여 penalize하지 못하는 단점을 가지고 있었다. 반면에 shift score는 예측된 alignment와 structural alignment와의 shift 정보를 바탕으로 하여, under-alignment나 over-alignment 뿐만 아니라 misalignment에 관해 1차원의 single number로 기술한다. 따라서 본 연구에서는 shift score가 alignment accuracy를 표현해주는 가장 좋은 척도라고 가정하였다. 본 논문에서는 Support Vector Regression (SVR) (Smola and Schölkopf, 2004) 에 의해 구조가 알려지지 않은 단백질에 대하여 shift score를 예측하는 새로운 방법에 대해 기술하도록 하겠다.

신경회로망(Neural Network; NN)이나 Support Vector Machine (SVM) 과 같은 기계학습방법 (machine learning technique)은 fold recognition 분야에서 매우 많이 쓰이는 방법이다. 그렇지만 이 방법들은 feature vector가 정해진 크기이어야 한다는 단점을 가지고 있다. 이러한 단점을 보완하기 위해 template library의 모든 template에 대하여 각각의 다른 길이의 profile-profile alignment 스코어를 갖는 feature vector가 최근 개발되었다(Han et al., 2005). 이 방법은 fold, superfamily, family 레벨에서 모두 향상된 결과를 보였다. 본 연구에서는 Han et al. (2005)에서 사용된 feature vector를 조금 수정한 것을 사용하였다.

재료 및 방법

Data

Template library를 만들기 위해서 SCOP version 1.69 (Murzin *et al.*, 1995)에 의한 분류가 사용되었다. 우선 ASTRAL Compendium (Chandonia *et al.*, 2004)에 의해 미리 준비되어진 40% sequence identity이하의 7400여 개의 도메인으로 구성되어진 fold library가 만들어졌다 (Chandonia *et al.*, 2004). 그 중에서 SVR의 training과 testing을 위해 적어도 10개 이상의 멤버를 가지고 있는 fold만을 선택하였다. 결과적으로 167개의 fold의 5190개의 template들이 선택되었다. 각 fold의 모든 template들의 2/3가 임의적으로 선택되었고 각 template들의 SVR을 training하기 위해 사용되었다. 나머지 1/3의 template들은 testing을 위해 사용되었다. 이러한 과정을

3-fold cross-validation을 위해 각각의 template들에 대하여 3번씩 이루어졌다.

Structural alignment

Alignment의 정확도를 정량화하기 위해, 즉 shift score를 계산하기 위해서 best alignment가 결정되어야 한다. 본 연구에서는 structural alignment가 best alignment라고 가정되었다. Structural alignment는 가장 좋은 alignment가 아닐 수도 있다. 하지만 구조가 알려진 단백질의 경우에는 현재의 존재하는 alignment 방법 중 가장 좋은 방법이다. 모든 template들에 대하여 fold 레벨에서 그들의 모든 이웃 단백질과의 structural alignment가 combinatorial extension (CE) algorithm (Shindyalov and Bourne, 1998)에 의해 만들어졌다. 주목하여야 할 점은 최종적으로는 structural alignment가 필요하지 않다는 것이다. 구조적인 정보를 이용하는 structural alignment는 SVR training을 위하여만 사용된다.

Profile-profile alignment and SVR feature vectors

모든 template의 training set에 대하여 SVR을 training하기 위해 Han *et al.* (2005)에 의해 개발된 방법을 수정하여 사용하였다. 일단 24가지의 다양한 alignment 옵션으로 어떠한 구조적 정보를 사용하지 않은 채로 all-against-all profile-profile alignment를 만든다. 다양한 옵션들은 gap open-penalty, gap extension-penalty, base-line score 등을 여러 가지 값으로 변화시켜서 만들었다. Query q의 i번째 위치와 template t의 j번째 위치가 정렬되었을 때의 profile-profile 스코어는 $m_{ij}=\sum_{k=1}^{20}(f_{ik}^{q}S_{jk}^{t}+S_{ik}^{q}f_{jk}^{t})/2$로 주어진다. 이 때 $f_{ik}^{q}, f_{jk}^{t}, S_{ik}^{q}, S_{jk}^{t}$는 query q의 i 번째 위치의 아미노산 k와 template의 j 번째 위치의 아미노산 i의 빈도수와 position-specific score matrix (PSSM) 스코어를 각각 나타낸다. 빈도수 행렬과 PSSM은 PSI-BLAST (Altschul et al., 1997)를 계산 횟수만을 바꾸고 (j = 6) 다른 옵션은 기본 옵션을 사용하여 만들었다.

길이가 n인 각각의 template들에 대하여 training set에 포함된 다른 template들과의 alignment가 만들어진다. 그 후에 이 alignment들이 각각 (*sa1*, *sa2*, ···, *sai*, ···, *san*, *total_score*, *query_length*) 의 형태를 가지는 (n+2) 차원의 feature vector로 변환된다. 이 때 *sai*는 주어진 template의 i 번째 위치에서의 profile-profile alignment 스코어를 나타내고, *total_score* 는 total profile-profile alignment score, *query_length* 는 template에 정렬된 query의 서열 길이를 나타낸다 (Fig. 1). 만약 gap이 alignment 상에 존재한다면 미리 정해진 임의의 음수 값을 부여한다. Alignment 스코어를 사용할 때에는 raw alignment 스코어를 사용하는 대신에 $sa^{i}=m_{i-2}+2m_{i-1}+3m_{i}+2m_{i+1}+m_{i+2}$로 계산되는 smoothened profile-profile alignment 스코어를 사용한다. m_i는 template의 i 번째 위치에서의 raw profile-profile alignment 스코어로 계산된다. (Tress *et al.*, 2003)

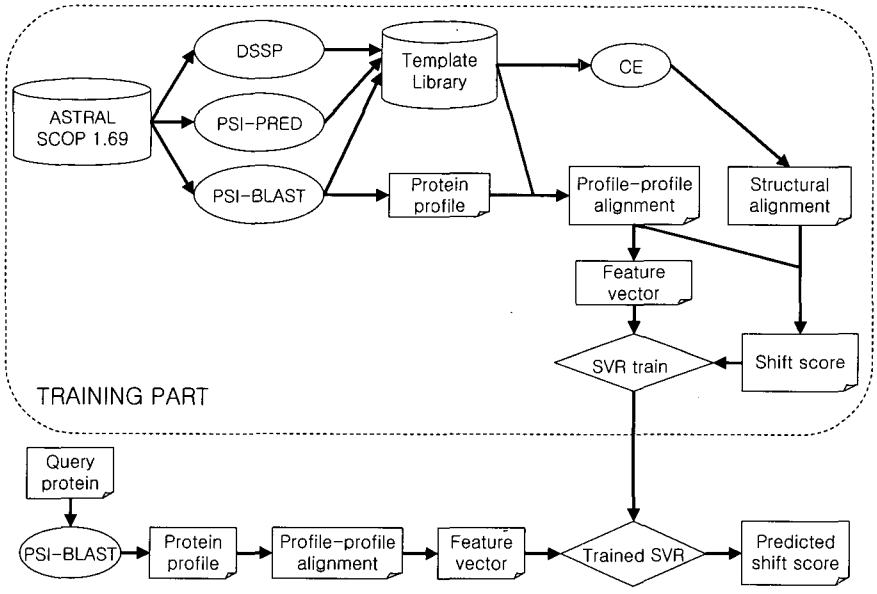

그림 2. SVR training과 testing까지의 개요. 점선의 안쪽 부분은 template 라이브러리의 구축, structural alignment, profile-profile alignment, shift score의 계산 과정을 포함하는 training 부분이다. 바깥쪽은 테스트하고 실제로 shift score를 예측하는 부분을 보여준다. 그림에서 보여지듯이 어떠한 구조적 정보도 실제로 alignment의 우수성을 예측하는 데에 사용되지 않는다.

Shift score (target of SVRs)

Alignment의 질을 나타내는 척도인 shift score가 본 연구의 SVR의 target으로 사용된다. 모든 structural alignment와 profile-profile alignment 쌍에 대하여 그들의 shift score가 ε =0.2으로 parameter를 설정하여 Cline et al. (2002)이 개발했던 방법과 같은 방법으로 계산되었다. Shift score는 - ε과 1.0 사이의 값을 가지므로 계산되는 shift score의 최소값은 - 0.2이다.

SVR training and testing and performance assessment

Training sample에 속하는 input-target 데이터에 대하여 각각의 SVR이 training 되었다. Training을 할 때 SVR의 kernel은 radial basis function (RBF) kernel을 사용하였고, parameter γ를 변화시켜가면서 SVMlight version 6.01 (Joachims, 1999)을 이용하여 training을 하였다. γ는 RBF kernel에 사용되는 parameter이다.

3-fold cross-validation을 수행하기 위해서 각각의 template에 대해 2개의 SVR이 존재하게 된다. 예를 들어 모든 template들이 set 1, set 2, set 3로 나뉘어진 경우에, 어느 특정 template가 set 1에 속한 경우 2개의 SVR 중 하나는 set 3를 테스트하기 위해서 그 template와 set 1과 set 2에 속하는 단백질들과의 alignment가 training example로 사용되고 다른 하나의 SVR은 set 2의 단백질들을 테스트하기 위해서 그 template와 set 1과 set 3에 속하는 단백질들과의 alignment가 training example로 사용된다. Fig. 2는 라이브러리의 구축과정에서부터 SVR의 training과 testing까지의 과정을 요약한다.

예측 정확도는 실제 shift score와 예측된 shift score 사이의 Pearson 상관계수로 평가된다. Pearson 상관계수는 $\frac{\sum_i (o_i - \bar{o})(p_i - \bar{p})}{\sqrt{\sum_i (o_i - \bar{o})^2 \sum_i (p_i - \bar{p})^2}}$로 주어진다. o_i는 실제 shift score, p_i는 예측된 shift score, $\bar{o}$ 와 $\bar{p}$는 각각 실제의 shift score, 예측된 shift score의 평균값을 의미한다.

표 1. 여러 training 옵션에 따른 Pearson 상관계수. SVR의 kernel 함수는 $\exp(-\gamma\|\mathbf{x}-\mathbf{y}\|^2)$로 주어진다. γ 값이 0.0001일 때, SVR이 shift score를 예측하는 데에 있어 가장 좋은 결과를 보여준다.

	γ=0.01	γ=0.001	γ=0.0001	γ=0.00001
Correlation	0.6845	0.7826	0.7960	0.7725

결과 및 고찰

SVR이 몇 개의 다른 alignment option에 대하여 Fig. 1과 Fig. 2에 묘사된 바와 같이 training된 후에 테스트되었다. 그 결과가 Table 1과 Fig. 3에 나타나있다. 결과를 평가하기 위해서 Pearson 상관계수가 사용되었다.

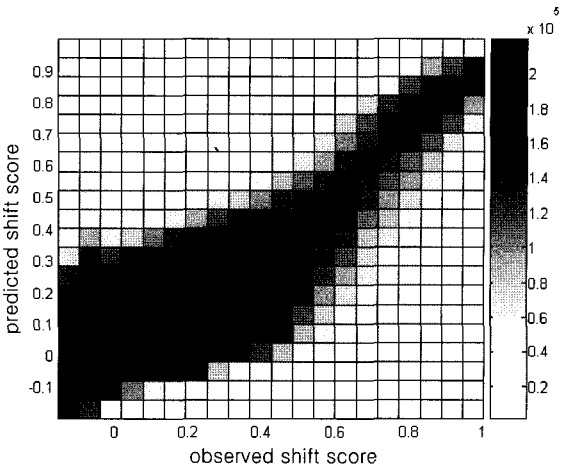

그림 3. 예측된 shift score와 실제 shift score와의 상관관계. 모양이 대각선에 가까울수록 예측이 더 정확하다.

Table 1은 서로 다른 parameter를 사용한 각각의 SVR에 대한 상관계수를 보여준다. 표에서 보여지듯이 γ가 0.0001인 경우에 가장 좋은 결과를 보인다는 것을 알 수 있다. 이 때 γ는 SVR의 RBF kernel $\exp(-\gamma\|\mathbf{x}-\mathbf{y}\|^2)$에 사용되는 parameter이다.

Pearson 상관계수 값이 0.7960을 갖는 γ=0.0001에 대하여 Fig. 3는 실제의 shift score와 예측된 shift score 사이의 상관관계를 보여준다. 그림에서 가장 높은 밀도를 갖는 부분은 검은색 정사각형으로 표시되고 가장 낮은 밀도를 갖는 부분은 하얀색 정사각형으로 표시된다. 전체적으로 큰 shift score를 갖는 점들보다 작은 값의 shift score값을 갖는 점들이 더 많다는 것을 알 수 있다.

예측된 값과 실제의 값 사이의 완벽한 상관관계의 경우에는 그림의 모양이 가는 대각선 형태를 띠게 된다. 본 연구의 경우에는 비교적 의미 있는 정도의 대각선 형태를 보인다. 이러한 의미 있는 상관계수와 대각선 모양의 밀도를 보이는 그림은 shift score가 본 연구에서 어느 정도 제대로 예측되었다는 것을 의미한다.

지금까지 structural alignment와 profile-profile alignment 사이의 shift score를 예측하는 방법은 없었다. 따라서 본 연구의 결과와 기존의 다른 연구 결과들의 직접적인 비교는 불가능하다. 하지만 최근에 alignment accuracy의 예측하는 접근방법을 fold recognition에 적용하려는 새로운 방법이 Xu (2005)에 의해 시도되었다. 그 방법은 0.72의 상관계수 값을 보였다. 다른 방법과 다른 데이터를 사용하였기 때문에 본

연구의 상관 계수 0.7960이 더 정확한 방법이라고 단정 지을 수는 없다. 하지만 regression에서의 상관계수 값의 비교를 통해 본 연구의 결과가 사용될 수 있을 정도로 regression이 되었다는 것과 어느 정도의 의미 있는 결과를 보였다는 것을 알 수 있다. 그리고 Xu의 방법이 fold recognition에서 성공적인 결과를 보였듯이 본 연구도 더 정확한 fold recognition에 활용될 수 있는 여지를 보여준다.

더 나아가 현재는 가장 좋은 alignment를 얻어내는 일반적인 옵션이 없기 때문에 이러한 방법이 shift score의 최대값을 찾아가는 최적화를 통하여 best alignment를 얻는 데에 활용할 수 있을 것이다.

요약하면 본 연구에서는 alignment의 질을 예측하기 위해 길이가 n인 template 에 대하여 $(n+2)$ 차원의 feature 벡터를 input 벡터로 사용하는 SVR로 shift score를 예측하는 새로운 방법을 개발하였고 테스트를 한 결과 의미 있는 결과를 보였다.

Shift score는 구조적 정보를 필요로 하는 structural alignment와 profile-profile alignment 사이의 차이점에 의해 결정되지만 본 연구에서의 방법은 두 alignment 사이의 shift score를 계산하는 데에 있어 어떠한 구조적 정보도 필요로 하지 않는다는 점이 주목되어야 할 것이다. 이 새로운 방법은 구조적 정보가 없이 가장 좋은 alignment를 찾는 데에 활용될 수 있을 것이다.

참고문헌

[1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, 25, 3389-3402.

[2] Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M. and Brenner, S.E. (2004) The ASTRAL Compendium in 2004, *Nucleic Acids Res*, 32, D189-192.

[3] Cline, M., Hughey, R. and Karplus, K. (2002) Predicting reliable regions in protein sequence alignments, *Bioinformatics*, 18, 306-314.

[4] Edgar, R.C. and Sjolander, K. (2004) A comparison of scoring functions for protein sequence profile alignment, *Bioinformatics*, 20, 1301-1308.

[5] Han, S., Lee, B.C., Yu, S.T., Jeong, C.S., Lee, S. and Kim, D. (2005) Fold recognition by combining profile-profile alignment and support vector machine, *Bioinformatics*, 21, 2667-2673.

[6] Holmes, I. and Durbin, R. (1998) Dynamic programming alignment accuracy, *J Comput Biol*, 5, 493-504.

[7] Joachims, T. (1999) Making large-scale support vector machine learning practical. In Schölkopf, B., Burges, C.J.C. and Smola, A.J. (eds), *Advances in kernel methods : support vector learning*. MIT Press, Cambridge, Mass., pp. 169-184.

[8] Jones, D.T. (1997) Progress in protein structure prediction, *Curr Opin Struct Biol*, 7, 377-387.

[9] Jones, D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences, *J Mol Biol*, 287, 797-815.

[10] Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L. and Hughey, R. (1999) Predicting protein structure using only sequence information, *Proteins*, Suppl 3, 121-125.

[11] Kim, D., Xu, D., Guo, J.T., Ellrott, K. and Xu, Y. (2003) PROSPECT II: protein structure prediction program for genome-scale applications, *Protein Eng*, 16; 641-650.

[12] Kinch, L.N., Wrabl, J.O., Krishna, S.S., Majumdar, I., Sadreyev, R.I., Qi, Y., Pei, J., Cheng, H. and Grishin, N.V. (2003) CASP5 assessment of fold recognition target predictions, *Proteins*, 53 Suppl 6, 395-409.

[13] Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F. and Sali, A. (2000) Comparative protein structure modeling of genes and genomes, *Annu Rev Biophys Biomol Struct*, 29, 291-325.

[14] Moult, J. (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction, *Curr Opin Struct Biol*, 15, 285-289.

[15] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol*, 247, 536-540.

[16] Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints, *J Mol Biol*, 234, 779-815.

[17] Sauder, J.M., Arthur, J.W. and Dunbrack, R.L., Jr. (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments, *Proteins*, 40, 6-22.

[18] Shindyalov, I.N. and Bourne, P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng*, 11, 739-747.

[19] Smola, A.J. and Schölkopf, B. (2004) A tutorial on support vector regression, *Statistics and computing*, 14, 199-222.

[20] Thompson, J.D., Plewniak, F. and Poch, O. (1999) A comprehensive comparison of multiple sequence align-

ment programs, *Nucleic Acids Res*, 27, 2682-2690.
[21] Tress, M.L., Jones, D. and Valencia, A. (2003) Predicting reliable regions in protein alignments from sequence profiles, *J Mol Biol*, 330, 705-718.
[22] Xu, J. (2005) Fold recognition by predicted alignment accuracy, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2, 157-165.
[23] Xu, Y. and Xu, D. (2000) Protein threading using PROSPECT: design and evaluation, *Proteins*, 40, 343-354.