GA-Based Inference of Euler Angles for Single Particle Analysis

Shusuke Saeki[1], Kiyoshi Asai[2], Katsutoshi Takahashi[2], Yutaka Ueno[2], Katsunori Isono[3], and Hitoshi Iba[1]

[1] Graduate School of Frontier Science, The University of Tokyo, Hongo 7-3-1 Bunkyo-ku, Tokyo, 113-8656, Japan
{saeki,iba}@miv.t.u-tokyo.ac.jp
[2] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Aomi 2-41-6, Koutou-ku, Tokyo 135-0064, Japan
{asai-cbrc,takahashi-k,yutaka.ueno}@aist.go.jp
[3] INTEC Web and Genome Informatics Corporation, 1-3-3 Shinsuna, Koto-ku, Tokyo 136-0075, Japan
isono@isl.intec.co.jp

Abstract. Single particle analysis is one of the methods for structural studies of protein and macromolecules developed in image analysis on electron microscopy. Reconstructing 3D structure from microscope images is not an easy analysis because of the low resolution of images and lack of the directional information of images in 3D structure. To improve the resolution, different projections are aligned, classified and averaged. Inferring the orientations of these images is so difficult that the task of reconstructing 3D structures depends upon the experience of researchers. But recently, a method to reconstruct 3D structures is automatically devised [6]. In this paper, we propose a new method for determining Euler angles of projections by applying Genetic Algorithms (i.e., GAs). We empirically show that the proposed approach has improved the previous one in terms of computational time and acquired precision.

1 Introduction

Structural analysis of proteins is currently conducted using primarily NMR and X-ray methods. However, each of these has limitations: NMR is applicable only to relatively small proteins, and X-ray analysis is constrained by the difficulty and limitations of crystallization.

Recently, a technique called single-particle analysis has been recognized as a viable method for analyzing three-dimensional protein structures that are difficult to analyze through other methods [2].Using this technique, a protein is frozen for observation by electron microscopy, and the three-dimensional structure is then determined from images of the protein in various orientations. This technique does not require crystallization of the protein and therefore very large proteins can be analyzed. However, the presence of noise in the raw micrographs causes the protein images to be barely recognizable, making resolution of the

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2724, pp. 2288–2300, 2003.

three-dimensional structure problematic due to the extremely small size of the proteins and their fragility under the electron beam.

In order to successfully obtain a protein image from noisy photomicrography, resolution must be enhanced by classifying multiple two-dimensional images according to orientation and then overlaying the images to determine the unknown three-dimensional structure [4]. This requires very complex image analysis techniques. Typically, a few tens of averaged images are obtained from 10,000 to 100,000 raw projection images, and from these, the three-dimensional structure is determined.

This method has finally become practical due to improvements in micrographic techniques and the recent rapid progress in computational capabilities [9,8]. In this paper, we propose a new algorithm based on a GA to address an especially difficult image analysis problem involved in the single-particle technique, i.e. obtaining a clear three-dimensional structural projection from images of proteins at various orientations.

2 Euler Angle Determination Problem in Single-Particle Analysis

2.1 Common-Line Method

Reconstructing the three-dimensional structures requires determination of the projection angle of each image. By obtaining three-dimensional positional relationships between images, practical techniques used for CT scan analysis can be applied to obtain a density distribution of the three-dimensional structure, which then allows the density distribution to be displayed by assigning a certain threshold value. In the three-dimensional density function, the particle orientation with respect to the image plane is represented by the Euler angle.As shown in Fig. 1, the Euler angle is composed of three fundamental angles:@U?¿,À and Á.

The basis of the Euler angle representation of the projection angle of each image is the common-line method [11,3], which holds well because micrographic images are transparent projections.

In the three-dimensional density function, the profile that is projected onto a straight line at a certain angle is common to two projection images, and can be obtained by projecting either projection image onto a straight line at a certain angle.

When there are three projection images, the projection angle of the three projection images is determined by the profile (common line) that matches the given angle.

Even when the common line for two projected images is determined, the three-dimensional projection direction cannot be expressed by a single set of parameters, as one rotational degree of freedom remains (Fig. 2a). At least three

projection images are required in order to express the three-dimensional projection direction using a single set of parameters (Fig. 2b) [5].

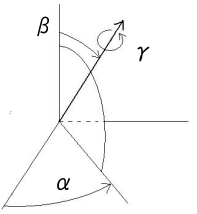

Fig. 1. Euler Angles ¿,À,Á

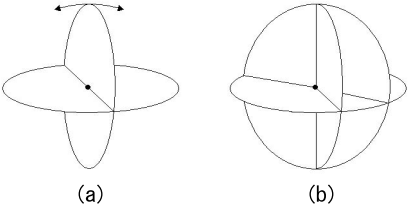

Fig. 2. Projections of a 3D Object.

2.2 Actual Common-Line Determination

Figure 3 illustrates the actual procedure used to determine the common line from sample projection images A and B. First, the profile, which is a three-dimensional density function called a sinogram, is obtained and then projected on a straight line at a certain angle (Fig. 3, left). The profile can be calculated to obtain the projection on the straight line while varying the angle of each of the projection images.

Next the sinogram of projection image A (angle θ_a) is compared to the sinogram of projection image B (angle θ_b). The index of similarity is expressed as the root mean square deviation (RMSD).By representing the sinograms of projections A and B as $snA(\theta_a, x)$, $snB(\theta_b, x)$, respectively, the RMSD value is expressed generally as follows:

$$RMSD(\theta_a, \theta_b) = \sqrt{\int_x \{snA(\theta_a, x) - snB(\theta_b, x)\}^2 \, dx}$$

The profile obtained by acquiring RMSD values by varying θ_a and θ_b, respectively, from 0 to 360 is called the cross sinogram, as shown on the right in Fig. 3.

2.3 Inference of the Euler Angle

If the Euler angles are given, RMSD values can be calculated.Therefore, the inference of the Euler angles based on common-line method boils down to the following optimization problem:

> **One of the projection images is regarded as a reference image and its Euler angle is fixed as (0, 0, 0).Determine the Euler angles of the rest of the projection images to minimize the sum of RMSD values of all projection image pairs.**

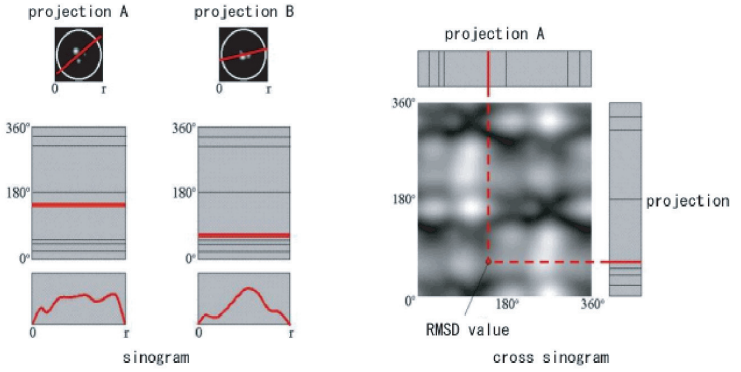

Fig. 3. Sinogram and Cross Sinogram.

Therefore, if N is the number of projection images, this optimization problem is a function of $3(N-1)$ parameters. Judging from the cross sinogram of Fig. 3, the sum total of functions of RMSD values has a fairly complicated shape, which we believe would result in a huge number of local solutions, implying the need for a highly efficient search method.

3 Search Method Using GA

3.1 Real-Coded GA

GA is a population-based search method by means of simulated evolution in a computer. The parameters of the optimization problem (the optimization problem function) are coded as genes, and the solution is searched by repeating the operations of selection, crossover and mutation.

Binary or gray representation is generally used for GA coding; however, for real-coded handling, a problem exists in that the phase structures of the genotype and the phenotype are too dissimilar. Thus, in recent years, the real-coded GA, which uses directly the actual number of vectors for coding, has been proposed. One example of the successful practical application of real-coded GA is the lens design system [7]. As the crossover method, Unimodal Normal Distribution Crossover (UNDX) is used in the current experiment. The UNDX generates offsprings around the line segment connecting two parents. It was shown that the UNDX can efficiently optimize some benchmark functions with strong epistasis among parameters compared to other crossover methods such as BLX-¿. Euler angles which we are going to optimize is considered to be the parameters with such a feature.

3.2 Sequential Additive Search Method

When applying the real-coded GA to optimize all Euler angles simultaneously, the precise three-dimensional reconstruction was not successful. Because the search space was simply too large, our investigations were steered toward one of the innumerable existing local solutions.

Therefore, consideration was given to introducing heuristics that would satisfy the problem.If a set of at least three images is provided, the optimized value among these images can be obtained. Thus, the following method of sequential additive optimization was considered.

1. Three images are optimized by real-coded GA. The Euler angle of the initial population is taken at random.
2. One image to be added is selected from among the remaining images.
3. The initial Euler angle of the original set of images is treated as the best solution of the previous optimization, and the Euler angle of the additional image generates a population as a random value.
4. Optimization is performed by real-coded GA.
5. Further optimization is performed by the steepest-descent method.
6. If an image to be added remains, return to step 2. If no images remains, finish.

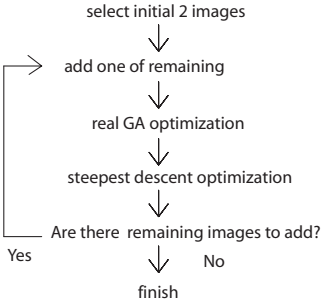

Fig. 4. Flow of Optimization

Figure 4 illustrates the flow of sequential additive search by GA. The best solution acquired by GA is further optimized by the steepest-descent method, because even though the precision remains unchanged, the required computation time is reduced to approximately one-tenth by halting the search at the proper generation. The best local solution is obtained by force from the steepest-descent method rather than searching through lengthy generations using GA.

Figure 5 shows the initialization details for this method. By including the best solution of the previous optimization as an initial individual, a more optimal initial population can be generated than for the case in which all parameters are initially random. Since the crossover method is UNDX [7], the best solution portion of the previous optimization of an individual can be changed, except by mutation. By this method, we believe that an incredibly large search space can be searched efficiently.

A problem associated with this method is that the quality of the optimization solution depends on the order of the additional images. The order of image addition can produce approximately a 10“ variation in RMSD value. Trying all possible orders is not realistic because the number of cases increases exponentially with the number of images. For the heuristic, a method that tries all possibilities of addition at each step while adding images one-by-one and then selects having the lowest RMSD value, is considered. However, with this additional method, because a GA is used for optimization, some unevenness remains in the quality of the solution at certain levels for each trial. To stabilize the

1. Randomly select two sets of images as the number of candidates.
2. Add one image from those remaining in each set, and perform optimization and verification. Check all additional possibilities.
3. From the new image set obtained by the above addition, select the number of candidates in sequence toward higher evaluation, and designate this as the next image set.
4. If an image to be added remains, return to step 2. If no images remain, finish.

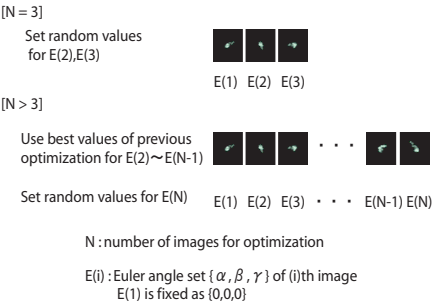

Fig. 5. Initialization of GA Optimization.

quality of the solution, a number of solutions are retained as "candidates" for the order of image addition during the search (Fig. 6).

1. Randomly select two sets of images as the number of candidates.
2. Add one of the remaining images to each set, and perform optimization and verification. Check all additional possibilities.
3. From the new image set obtained by the addition, select the number of candidates in sequence toward higher evaluation, and designate this as the next image set.
4. If an image to be added remains, return to step 2. If no image remain, finish.

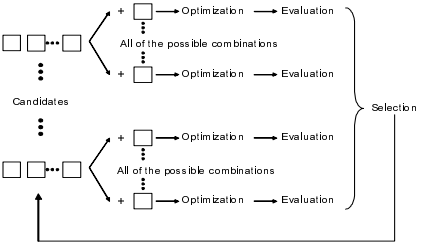

Fig. 6. Flow of Image Addition.

4 Experimental Results with the Three-Dimensional Structures Reconstructed from Projection Images

4.1 Experimental Projection Images

In order to evaluate the method by GA, an experiment was performed using a protein having a known three-dimensional structure. The protein used for

this experiment was myosin, which generates power in muscle tissue. Figure 7 shows the three-dimensional structure of myosin and twenty images acquired by projection from different directions. Each pixel of the image takes a real value from 0.0 to 256.0. The objective was to determine the Euler angles of these 20 projection images.

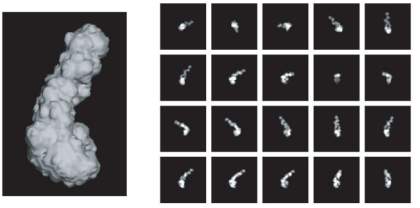

Fig. 7. Projection of Myosin.

4.2 Validity of Sequential Additive Search

Reconstruction of the three-dimensional structure was performed using the GA-based method described earlier. The image size is 65 x 65 pixels. In order to investigate the validity of sequential additive search by GA, comparison of a result with global search by GA(i.e. real-coded GA holding a vector of 57 variables from the start) was performed. The parameters for global search by GA are shown listed in Table 1. The relationship between the number of generations and the RMSD values in global search by GA is shown in Fig. 8. The best search, the worst search, and the average in ten trial are displayed. The last result was further optimized by the steepest descent method. RMSD values were saturated around 1.0 in a certain amount of generation.

Table 1. The Parameters for Global Search.

Population	100
Crossover	0.8
Mutation	0.00175
Generation	15000

Table 2. The Parameters for Sequential Additive Search(N : number of images).

Population	50
Crossover	0.8
Mutation	$0.3/\{3(N-1)\}$
Generation	25N
Repeat	5

The parameters for sequential additive search by GA are listed in Table 2. The number of mutations was selected to be proportional to the gene length (i.e., the number of parameters). In addition, the number of generations was selected so as

to be proportional to the gene length. The search process was repeated five times, and the best solution among the obtained solutions was selected, thus scattering of the solution was minimized. We found that the solution quality tended to be better by performing five trials (resulting in five times the population) than by making the number of individuals in the population five times larger, although the computational cost was approximately the same. Because the population approaches uniformity with the evolution of individuals, we believe that various evolved individuals can be ultimately obtained if the population number is larger.

A result of seqential additive search is shown in Figs. 9, 10, and 11. The search by GA was tried 10 times to each number of candidates, in order to investigate the stability of the solution. Figure 9 plots the best value, the average value and the worst value to the number of candidates. We found that average value of RMSD values decreases when the number of candidates increased. Moreover, it turned out that the best value of RMSD values also decreases slightly to the increase in the number of candidates. Variance of the RMSD values to the number of candidates is shown in Fig. 11. The tendency for the variance of the solution to decrease exponentially to the number of candidates was observed.

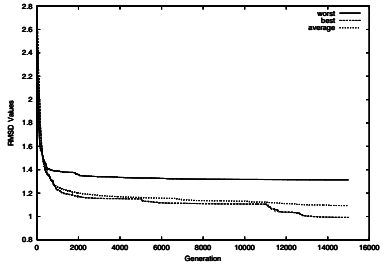

Fig. 8. RMSD Values v.s. Generations.

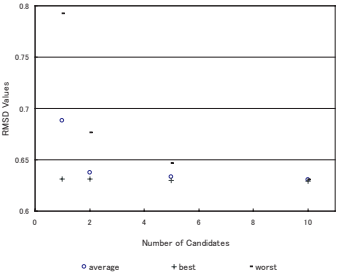

Fig. 9. RMSD Values v.s. Number of Candidates.

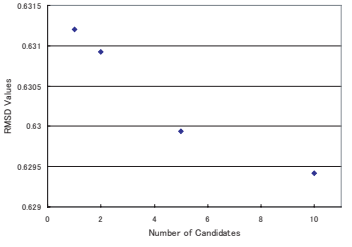

Fig. 10. Best of RMSD Values

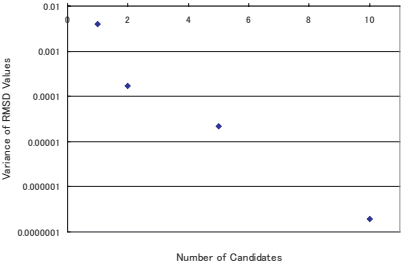

Fig. 11. Variance v.s. Number of Candidates.

Figure 12 shows the reconstructed three-dimensional structure obtained using the derived Euler angles. Figure 12(a) is a solid reconstruction based on the projection direction when the projection image was generated, so the purpose of this experiment was to reconstruct this structure. Figure 12(b) is a result of sequential additive search by GA with ten candidates of the solutions. Figure 12c is a result of global search by GA. Figure 12b is obviously closer to the target structure Fig. 12a as compared with Fig. 12c. When the result of global search is 0.92 in a RMSD value, the result of sequential additive search is 0.63 in a RMSD value. It can be said that sequential additive search is superior to global search in this problem.

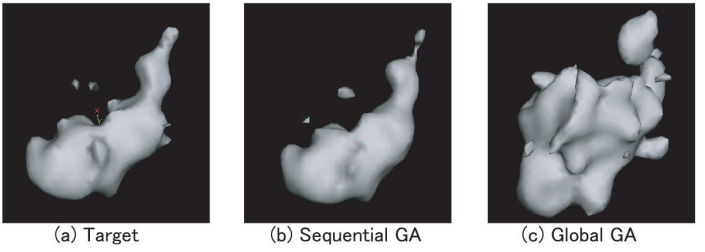

Fig. 12. Reconstructed 3D Structures.

4.3 Effect of Noise and Blur

In order to investigate the applicability to real micrographic images, some noise and blur were added and their effects were examined. Gaussian noise was applied, and blur was added using a Gaussian filter. Gaussian noise is a model of real noise in raw micrographic images, whereas Gaussian filter can simulate blur phenomena caused by the clustering operations. Figure 13 shows the images with noise and blur. The level in Fig. 13 is the Gaussian noise average level parameter, and half of a standard deviation was added. The noise factor and intensity follow a Gaussian distribution. SN in Fig. 13 is the parameter of the Gaussian filter representing the standard deviation of the angle at which pixels are blurred on a circle centered on the image. When SN is zero, no filtering is applied. The Gaussian noise was added first, followed by the Gaussian filter.

Reconstruction was performed by sequential additive search by GA with 50 candidates. The parameters for GA are the same as Table 2. The image size used this experiment is 96 x 96 pixels in order to show that our proposed method is applicable also to the image of other resolution. Figure 15 shows the target the three-dimensional structure reconstructed using the correct Euler angles. The arrangement of the three-dimensional structure corresponds to that of the projection image in Fig. 13. Figure 14 is a solid reconstruction based

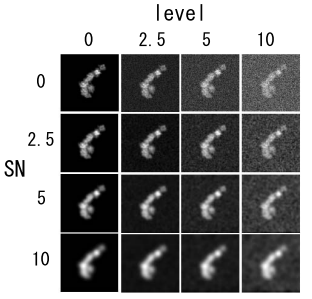

Fig. 13. Projections with Gaussian Noise and Gaussian Filter

on the projection direction when the projection image was generated, so the purpose of this experiment was to reconstruct this structure. When no noise was added, the reconstruction was almost perfect. We observed a tendency for the reconstructed three-dimensional structure to be smoothed when the Gaussian filter was applied. However, when the Gaussian noise level was set to ten, we were still able to clearly discern the tail structure of myosin when using the Gaussian filter. We believe that the influence of Gaussian noise was reduced by the blurring. It could be claimed that a certain degree of degradation of the images is a valid sacrifice for this noise reduction when the global three-dimensional structure is of interest.

Fig. 14. Target 3D Structure.

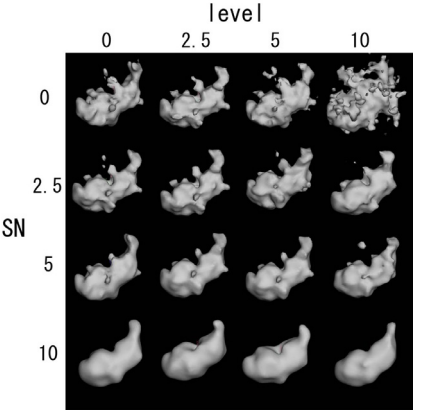

Fig. 15. Reconstructed 3D Structures.

Table 3. Comparison with Other Methods.

	RMSD for inference	RMSD with target 3D structure	Time(hour)
(a)sequential GA	0.03136	5.11	30
(b)spot detection	0.03160	5.12	110
(c)global GA	0.04358	11.35	30

4.4 Comparison with the Previously Proposed Method

There was previously proposed method called spot detection method. This method is based on a sequential additive search method identical to the present method, except that the solution candidate is searched by detecting a comparatively large area (spot) of mutual correlation of cross sinograms [12,10]. For added images, the ¿ and Á components of the Euler angle are obtained from the spot information of the cross sinogram between the standard image and the added image, and À is estimated from the overall relationship.

Table 3 lists the results for the experimental RMSD values and the required computation times for this method along with those of various other methods. The microprocessor used for the present experiments was an Athlon XP1800+. Table 3(a) shows the case using fifty candidates in the present method. Table 3(b) is, as far as we know, the only method proposed in the past to solve this problem.Table 3(c) shows the results of searching a wide space via GA. Projection images with noise (level 10, sn 10) are used because acutual microscopic images contain a great deal of noise. RMSD values of 3D structures between the target and the acquired ones by these methods are compared to investigate the solution precision. The proposed method(a) is superior with respect to both computation time and solution precision, as compared to the spot searching method(b). Compared to these two methods, the wide-space search via GA(c) showed greatly reduced precision.

5 Discussion

Only the spot-detection method has been known as the single particle analysis so far, for the sake of inferring the Euler angles solely from the projection images, i.e., not by using the symmetrical information of 3D structures. The proposed GA-based method gave better performance to the noisy projection images, in terms of the computational time and the acquired precision. One reason seems to be that our approach is relatively independent from the reference images so that it can explore the search space more widely. On the other hand, in case of noise-free projection images,the spot detection method may be superior, especially when we handle the higher resolution. This is because the spot information can provide a more correct common line with the higher resolution. When applying to the real data, we inevitably have to cope with the noisy projection, which will make our method more practical and effective.

Our GA-based method is a stochastic search so that different solutions are acquired from one run to another.However, the RMSD values by our method were much smaller, i.e.,better,for all the runs. In general, many parameters (e.g., see Table 2) have to be effectively tuned for the sake of efficient GA-base search.We have tried two types of resolutions, i.e.,65x65 pixels and 96x96 pixels, for the projected images.As a result of several comparative experiments, we have confirmed that only the number of candidates should be changed to give the satisfactory performance to these cases.In other words, most of the other GA parameters need not be considered for the further tuning to cope with different experimental set-ups.

Our future reseach concerns are on reconstructing three-dimensional structures from projection images acquired by clustering operations.Clustering and averaging algorithms for the robust image-processing have been worked on [1]. We are currently working on the application to clustered and averaged projection images of a simulation and we plan to use actual micrographic images.

6 Conclusion

We found that three-dimensional structures could be restored from projection images by the method proposed in the present paper. In addition, noise was added to the projection images, and its effect on the reconstructed solid body was investigated. In processing, projection images that are judged to be oriented in the same direction are clustered and then overlaid to reduce noise by sacrificing image clarity through fading to a certain level. The experimental results suggest that the proposed type of processing is viable for discerning large target structures.

The present method was found to be superior compared to previous methods when noise is included in the images. Because actual micrographic images contain a great deal of noise, the present method is considered to be more practicable than previous methods. Introducing heuristics that correspond to the characteristics of the problems involved in searching via GA narrows the search space. Conversely however, for difficult problems, the present method can still search the space effectively, and thus is quite useful.

In the above described experiments, we restored a three-dimensional structure directly, based on projection images of a simulation having added noise. However, we must first investigate reconstruction of three-dimensional structures based on projection images acquired by clustering operations. Ultimately, we intend to apply the present method to actual micrographic images.

References

1. K.Asai, Y.Ueno, C.Sato K.Takahashi "Clustering and Averaging of Images in Single-Particle Analysis" Genome Informatics 11, 151–160 2000.
2. J.Frank "Three-dimentional Electron Microscopy of Macromolecular Assemblies" Academic Press, London 1996.

3. G. Harauz "Exact filters for general geometry three dimensional reconstruction" Optik 73, 146–156 1986.
4. van Heel, et al "Multivariate statistical classification of noisy images" Ultramicroscopy 13, 165–183 1984.
5. van Heel, et al "Angular reconstruction : a posteriori assignment of projection directions for 3D reconstruction" Ultramicroscopy 38, 241–251 1987.
6. van Heel, et al. "A new generation of the IMAGIC image processing system" J.Struct. Biol. 116, 17–24 1996.
7. I.Ono, S.Kobayashi "A Real-coded Genetic Algorithm for Function Optimization Using Unimodal Normal Distribution Crossover, Proc. of 7th Int. Conf. on Genetic Algorithms" Proc. of 7th Int. Conf. on Genetic Algorithms 246–253 1997.
8. S.Saeki, K.Asai, K.Takahashi, Y.Ueno, K.Isono, H.Iba "Inference of Euler Angles for Single Particle Analysis by Using Genetic Algorithms" Genome Informatics 12, 151–160 2002.
9. C.Sato, Y.Ueno, K.Asai, K.Takahashi, M.Sato, A.Engel, Y.Fujiyoshi "The voltage-sensitive sodium channel is a bell-shaped molecule with several cavities" *Nature*, Vol.409, No.6823, pp. 1047–1051, 2001.
10. K.Takahashi, Y.Ueno, K.Asai "Euler Angle Decision Method between Transfer Images of a Protein Molecule for Single Particle Analysis" The Biophysical Society of Japan, 39th Annual Meeting, 1P041, 2001.(in Japanese)
11. Penczek, et al. "Three-dimentional reconstruction of single particles embedded in ice" Ultramicroscopy 40, 33–53 1992.
12. Y.Ueno, K.Takahashi, K.Asai, C.Sato "BESPA: software tool for Three-Dimensional Structure Reconstruction from Single Particle Images of Proteins" Genome Informatics 10, 241–242 1999.