the calculation were improved by integrating these four new techniques. In the session, applications of our system will also be presented.

1E1348 ヒトプロテオームに存在する新規ホルド数の見積もり

An estimation of the number of new folds in the human proteome

Satoshi Fukuchi¹, Keiichi Homma², Ken Nishikawa² (¹*Maebashi Institute of Technology*, ²*National Institute of Genetics*)

Eukaryotic proteins consist of structural domains (SDs) and intrinsically disordered regions (IDRs), which have become a consensus based on the recent progress on the study of ID proteins. We developed a system, DICHOT, which classifies the entire protein sequence into SDs and IDRs. DICHOT produces two types of SDs, the known domains (KDs) and the cryptic domains (CDs), the latter have a potential to take a 3D structure but no information on 3D structures yet. In the application of DICHOT to the human proteome, the residue-wide percentage of KDs, CDs, and ID regions are 52%, 13%, and 35%. In the assigned KDs, we found 943 SCOP superfamilies. An attempt to cluster CD regions could not converge into a plausible number of families. Then, we approached it from another angle, using the statistics of Pfam domains found in KDs. Assuming that the ratios of the number of SCOP folds to the number of Pfam domains are constant in KD and CD, we obtained 519 SCOP folds in CDs. This figure is considerably larger than the one estimated under an assumption that the number of folds per residue is constant in KD and CD. This discrepancy may be explained by the following notion: KDs contain many ubiquitous folds found frequently in different proteins and cover a large fraction of regions in a proteome. By contrast, the structures in CDs are likely to contain rare folds, which need more folds to cover the same regions as ubiquitous fold classes do.

1E1412 Development of a method for predicting carbohydrate-binding proteins

Zhenyi Ge, Wei Cao, Kazuya Sumikoshi, Shugo Nakamura, Kentaro Shimizu (*Graduate School of Agricultural and Life Sciences, The University of Tokyo*)

Carbohydrate-binding proteins are important complexes that participate in many biological process as intercellular information transfer, cell recognition and cellular immunity. Carbohydrate-binding proteins are defined as carbohydrate-binding complexes which are not directly modified by carbohydrate. A number of eligible carbohydrate-binding proteins are searched from Uniprot Knowledgebase by SRS (Sequence Retrieval System). After refined, we constructed a positive data set and a negative data set whose sequences are randomly selected from the same database. SVM (Support Vector Machine) is employed to train and predict. SVM is a supervised learning method that can be applied to classification or regression. We used SVM to classify carbohydrate-binding proteins according to the types of binding carbohydrates.

1E1424 タンパク質立体構造モデリングの精度を上げる —アミノ酸残基の挿入欠失頻度の表面露出度による定量化—

A method to improve homology modeling of proteins - Insertion-deletion frequency depending on surface accessibility -

Atsushi Hijikata¹, Kei Yura², Mitiko Go^{3,4,5} (¹*RIKEN RCAI*, ²*Grad. Sch. Hum. Sci., Ochanomizu Univ.*, ³*Facul. Biosci., Nagahama Inst. of Bio-Sci. Tech.*, ⁴*Res. Org. Info. Sys.*, ⁵*Med. Res. Inst., Tokyo Med. Dent. Univ.*)

In homology modeling, the model quality largely depends on the quality of the sequence alignment between the target and template proteins. Although many efforts have dedicated to improve the alignment quality, there is no best single method for the alignment in cases where the target and template sequences share low sequence identity. Here we focused on better locations of insertions/deletions (gaps) in sequence alignments between distant homologous proteins. Revisiting gap locations on the three-dimensional structures of ~18,000 homologous protein pairs revealed that the frequency of gap locations approximated to an exponential function of the solvent accessibility. By introducing this relationship into the gap penalty function for pair-wise alignment between template and target sequences, we demonstrated that the quality of the alignments was substantially improved on a pair of sequences with identity in the "twilight zone" between 20 and 40%, especially for locating gap positions, and named this method ALAdGAP (ALignment with Accessibility dependent GAP Penalty; Hijikata et al., *Proteins*, 2011; http://cib.cf.ocha.ac.jp/target_protein/). In combination with a conventional method for model building we applied ALAdGAP for the template-based modeling in the CASP9

experiment, and found that the quality of models we submitted were as good as the best groups if the best template structure was selected accordingly, indicating that ALAdGAP is considerably suitable for the target-template sequence alignments in homology modeling.

1E1436 エキソン境界と構造アラインメントの関連が示唆する蛋白質多様化のメカニズム

A related location of exon boundary and structural alignment boundary reveals a mechanism of protein fold diversity

Shintaro Minami, Kengo Sawada, George Chikenji (*Grad. Sch. of Eng. Univ. of Nagoya*)

It is revealed that the sequence-order independent tertiary structural similarity between a lot of proteins. One of the most famous examples is a circular permutation that can be interpreted by a typical genetic process. However many others have more complex alignment that are hard to explain in existing genetic process. Generally, such pairs were considered to be structural analogous. In recent years, however, a protein pair that have a strong functional site similarity was discovered by Alex EJ and Valentin AA (2007) though it has the complex non-sequential structural similarity. In addition, the pair has been classified into same super-family in SCOP database means they have a common evolutionary origin. Now, some similar cases have been found by other groups, but it is still unclear how such protein have evolved. We performed here a comprehensive large-scale study of non-sequential structure comparison between SCOP family representative structures. As a results, many non-sequential pairs which have strong structural similarity were found. Furthermore, some of them also have considerable similarity in their sequence profile. To understand their evolutionary relationships, we identified positions of intron in the pair of amino acid sequences and compared them to structural alignment results. Then we found that introns often be found near non-sequential alignment boundary. We finally concluded that exon/intron boundaries are strongly related to structural diversity of protein domains in evolution.

1E1448 ノンローカルな相互作用を重視した新規タンパク質構造予測手法の開発

Development of a novel protein structure prediction method emphasizing non-local interaction

Kota Nakamori, Shintaro Minami, Kengo Sawada, George Chikenji (*Grad. Sch. of Engineering, Nagoya Univ.*)

Protein structure prediction is one of the most important and fundamental problems of biology. In particular, the prediction of novel protein structures, for which we cannot use template, is a very significant from the viewpoint of biophysics. Currently, one of the most successful method for predicting novel protein structures is the fragment assembly (FA) method. The success of the method is, however, limited to low contact order (CO) proteins, probably because the FA method emphasizes local interaction too much. (R.Bonneau. *Protein Sci.* 2002;11:1937-1944).

In order to predict novel protein structures with high CO, we have been developing a new method for novel protein structures prediction which is based on the opposite approach of the FA method, the method emphasizing non-local interaction. The method consists of the following two steps; First, we collect structural candidates for every pair of secondary structure elements (SSE) of the target sequence, retrieving them from the structural database by threading technique. In this stage, we also consider a reverse mapping of SSE. Second, we construct tertiary structures by assembling structures of pairs of SSE, searching for low energy structures. Preliminary results shows that our method is more effective for predicting high CO protein structures than the FA.

1E1512 二次構造予測に基づく束縛付きマルチカノニカル分子動力学シミュレーションによるタンパク質の立体構造予測

Protein tertiary structure prediction by multicanonical MD simulation with restraints based on secondary structure prediction

Asahiko Nishimura¹, Tohru Terada², Shugo Nakamura¹, Kentaro Shimizu^{1,2} (¹*Dept. of Biotechnology, Graduate School of Agricultural and Life Sciences, Univ. of Tokyo*, ²*Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, Univ. of Tokyo*)

We are developing a method of predicting the native structure of a protein using molecular dynamics simulation with restraints on the secondary structures. In

our previous study, we found that given correct secondary structure information, the native structure of a protein can be predicted very accurately [Furuta *et al.*, *Chem. Phys. Lett.* **472**, 134 (2009)]. However, because the correct secondary structures are not known a priori, it is necessary to obtain the information from secondary structure prediction. Since the prediction contains errors, we extended our method so that it can use the result of the uncertain secondary structure prediction. We used the PSIPRED program to predict the secondary structures. Restraining potentials were applied the backbone dihedral angles (ϕ , ψ) of each residue whose secondary structure is predicted to be α -helix or β -strand. The potential energy function was defined as follows. When the ϕ and ψ dihedral angles are within the predefined free areas ($-180^\circ \leq \phi \leq 0^\circ$ and $-90^\circ \leq \psi \leq 90^\circ$ for α -helices and $-180^\circ \leq \phi \leq 180^\circ$ and $-180^\circ \leq \psi \leq -90^\circ$ or $90^\circ \leq \psi \leq 180^\circ$ for β -strands), the potential is zero. When they go out of the areas, penalties are imposed on them. The height of the restraining potential was determined so that the probability of existence in the free area accords with the accuracy derived from the confidence score of the prediction. We applied this method to proteins with different folds. We will report the results and will discuss the effectiveness of this method.

1E1524 ドメインドメインドッキングに適したスコア関数の開発

Development of a scoring function for domain-domain docking to predict multi-domain protein structures

Satoru Hirako¹, Masafumi Shionyu² (¹*Grad. Sch. Bio-Sci., Nagahama Inst. Bio-Sci. Tech.*, ²*Fac. Bio-Sci., Nagahama Inst. Bio-Sci. Tech.*)

In prokaryotic and eukaryotic genomes, more than half of genes encodes multi-domain proteins. To elucidate the details of molecular function of these multi-domain proteins, understanding of relative spatial arrangement between each domain is important. However, the number of multi-domain proteins whose whole 3D structures are determined is small. Therefore, we have developed a structure prediction method for two-domain proteins based on domain-domain docking. To predict 3D structure of two-domain protein, we generate some candidate model structures from individual domains by ZDOCK program, and select the most plausible ones using a scoring function we have developed. The scoring function is a linear sum of docking score, domain interface score and terminal distance score. The docking score is obtained from ZRANK. The domain interface score is calculated from the number of predicted domain-domain interface residues used in the domain-domain interface of each candidate structure. To predict domain-domain interface residues, we have developed K(IP)2 method. The terminal distance score is calculated using the deviation from the estimated distance between the C-terminus of the first domain and the N-terminus of the second domain, which is obtained from the number of residues in the domain linker, in each candidate structure. We evaluated the performance of this scoring function on test sets. As a result, we confirmed that the scoring function could select more accurate models as compared to conventional docking methods.

1E1536 相互作用プロファイルを用いたタンパク質間ドッキングの代表相互作用の探索

Representative Interaction FingerPrint from decoys generated by rigid-body protein-protein docking

Nobuyuki Uchikoga¹, Yuri Matsuzaki¹, Masahito Ohue¹, Takatsugu Hirokawa², Yutaka Akiyama¹ (¹*Grad. Sch. Comput. Sci., Tokyo Inst. Tech.*, ²*CBRC, AIST*)

In protein-protein Interaction network, it is important to predict which pairs of proteins interact in living cell and how these bind each other. We approach to these problems using rigid-body docking algorithm, generating many decoys including false positives. However, in some cases, there are no near-native decoys by rigid-body docking. Then, we applied cluster analysis to generating representative decoys or interaction fingerprints rather than searching near-native decoys.

We have been developed Interaction FingerPrints (IFP) for the post-docking analysis of protein-protein rigid-body docking. On the other hand, although RMSD is one of the methods for calculating similarities between decoys, this value depend on the algorithm of superposing their three dimension structures. IFP is composed of binary states of interacting amino acid residues of each proteins, as a scale for measuring unique similarities between the complex structures. Then, IFP can be used easily to evaluate similarities between decoys. Additionally, IFP can be useful to assemble interaction profiles from various decoys.

In this work, we developed a method to generate Representative IFP (RepIFP) for rigid-body docking decoys. As results, we could obtain near-native RepIFPs,

which have the lowest energy scores among each pairs in 19 out of 44 protein pairs. And then, near-native RepIFPs were found in 2 out of 5 no near-native decoys cases. These results indicate that the RepIFP method is useful to predict protein-protein interactions in difficult cases.

1E1548 蛋白質複合体のための構造比較法 SCPC の開発

SCPC: A method to structurally compare protein complexes

Ryotaro Koike, Motonori Ota (*Grad. Sch. Info. Sci., Nagoya Univ.*)

Protein-protein interactions are of vital importance for various biological phenomena, and thus have been extensively explored. Physical contacts among proteins were revealed at the atomic resolution through three-dimensional structures of protein complexes determined experimentally. To examine a huge amount of protein complex structures in Protein Data Bank, an automatic method to compare them is mandatory.

We have developed SCPC, a novel method to structurally compare protein complexes. SCPC compares spatial arrangements of subunits in a complex with those in another complex using secondary structure elements. Similar sub-structures are detected in two protein complexes, and the similarity is scored. SCPC was applied to hetero dimers, homo oligomers and haemoglobins. Structural similarities of hetero dimers were properly estimated, showing good agreement with the results based on a pre-existing score. Conserved similar sub-structures were detected in a homo-tetramer and a homo-hexamer, both composed of homologous proteins. Classification of quaternary structures of haemoglobins was consistent with the conventional one. These results demonstrate that SCPC is a significant tool to investigate protein complex structures.

1E1612 空間統計学による立体構造情報を利用した機能部位予測のための相同配列選択手法

A method to select homologous sequences appropriate for functional region prediction by spatial statistics with structure information

Wataru Nemoto, Hiroyuki Toh (*AIST, CBRC*)

Cluster detection of conserved residues on a structure is one of the popular strategies for functional region prediction of the protein. Comparisons of homologous sequences are often utilized to evaluate the residue conservation. To distinguish between conserved residues and unconserved ones, a multiple sequence alignment of homologous sequences is constructed. An unsolved problem in such approaches is the selection of appropriate homologous sequences for the alignment. It is empirically known that a certain degree of sequence divergence in the alignment is essential for accurate predictions. However, there is no criterion for the divergence that is applicable to all cases. We have developed an index to select appropriate homologous sequences for the alignment by integrating structure and sequence information by spatial statistics. The index is designed to quantify the contrast of residue conservation on the molecular surface under the assumption that the prediction accuracy is the maximum when the contrast between functional regions and non-functional ones is the maximum. The prediction performances by the index are higher than those by other sequence-based methods. This implies that the integration of structure and sequence information to the index would be useful for the appropriate sequence selection. As far as we know, this is the first time that spatial statistics has been applied to protein analyses. Such integration of structure and sequence information by spatial statistics would be useful for other bioinformatics problems.

1E1624 活性部位とリガンド結合部位の情報を用いた酵素の機能予測

Use of information about active sites and ligand-binding sites for the prediction of enzyme functions

Chioko Nagao, Kenji Mizuguchi (*National Institute of Biomedical Innovation*)

We have analyzed systematically the active sites and ligand-binding sites of all enzymes in the Enzyme Catalytic-mechanism Database (EzCatDB) [1] and found a good correlation between known functional subclasses and the groups obtained by the clustering of the subsequences consisting of the active sites and ligand-binding sites [2]. In this study, we examined how such information can be utilized for the prediction of detailed enzyme functions. We compiled information about active sites and ligand-binding sites from all the CATH [3] domains, to which Enzyme Commission (EC) numbers were assigned. The active sites and ligands information were obtained from EzCatDB, Catalytic Site Atlas (CSA) [4] and PROCNATE [5] databases. The ligand binding