COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

Volume No: 8, Issue: 11, e201308007, http://dx.doi.org/10.5936/csbj.201308007

CSBJ

Recent advances in functional region prediction by using structural and evolutionary information — Remaining problems and future extensions

Wataru Nemoto [a,*], Akira Saito [a], Hayato Oikawa [a]

Abstract: Structural genomics projects have solved many new structures with unknown functions. One strategy to investigate the function of a structure is to computationally find the functionally important residues or regions on it. Therefore, the development of functional region prediction methods has become an important research subject. An effective approach is to use a method employing structural and evolutionary information, such as the evolutionary trace (ET) method. ET ranks the residues of a protein structure by calculating the scores for relative evolutionary importance, and locates functionally important sites by identifying spatial clusters of highly ranked residues. After ET was developed, numerous ET-like methods were subsequently reported, and many of them are in practical use, although they require certain conditions. In this mini review, we first introduce the remaining problems and the recent improvements in the methods using structural and evolutionary information. We then summarize the recent developments of the methods. Finally, we conclude by describing possible extensions of the evolution- and structure-based methods.

Mini Review Article

Introduction

The prediction of functional regions in a protein is an important research focus, and many methods have been developed for this purpose [1]. One of the most effective strategies is the detection of evolutionarily important residues on the tertiary structure of a protein, by integrating the structural and evolutionary information encoded in a multiple sequence alignment (MSA) [2–9] (see a schematic image of the strategy in Figure 1). The most popular and pioneering method based on the strategy is Evolutionary Trace (ET) [2], which uses a phylogenetic tree to rank the residues in a protein by their evolutionary importance and maps them on a closely related structure. The highly ranked residues are often clustered in space, and thus these clusters correspond to functionally important residues and are used to identify them. Many servers perform ET [10–12] or similar methods [3,5,7,13], and were developed by the original designers of ET [10] or other groups [3,5,7,11–13]. In this mini review, we will summarize the recent advances in the ET and ET-related methods (evolution and structure information-based methods) using structural and evolutionary information, including our work, over the past few years, and then discuss the remaining problems. First, we will summarize the various improvements of the measurements to evaluate the evolutionary information calculated from an MSA. We will subsequently introduce several improvements of functional region prediction by exploiting the structural information. We will finally introduce an important problem shared by the MSA-based methods in structural bioinformatics, and the challenges to solve it. At the end of this review, we will explain the potential extensions of the structure- and evolution-based methods. The web servers of the introduced methods and their update statuses are summarized in Table 1.

Recent advances

Improvements in the methods to evaluate evolutionary information

One of the most widely used scores to consider evolutionary information is the residue conservation at a site in an MSA. The residue conservation reflects the evolutionary selection at functional sites to maintain protein function and to retain structural folds [6], regardless of the developed conservation score formulae [14]. Therefore, the discrimination between the functionally important residues and the structurally important ones is often difficult [6]. This problem has led to limitations of the methods to predict the functional regions using conservation scores. In order to distinguish between the residues conserved for functional reasons and those conserved for structural constraints, Chelliah *et al.* [6] developed Crescendo. This program calculates the conservation scores with an Environment-Specific Substitution Table (ESST) [15], which describes the patterns of substitutions in terms of the amino acid locations within secondary structure elements, as well as the solvent accessibility and the existence of hydrogen bonds between side chains and neighboring residues. Crescendo [6] predicts functional regions by identifying clusters of residues with unusually high evolutionary restraints. To this end, they identified the evolutionary restraint at a site, as follows: 1) whether there is a high degree of evolutionary conservation than expected, 2) whether ESST makes weak predictions of the substitution patterns, and 3) whether there are residues within spatially conserved regions, when protein structures within the same

[a]*Division of Life Science and Engineering, School of Science and Engineering, Tokyo Denki University (TDU), Ishizaka, Hatoyama-cho, Hiki-gun, Saitama, 350-0394, Japan*

* Corresponding author. Tel.: +81 492961014
E-mail address: watarunemoto@gmail.com (Wataru Nemoto)

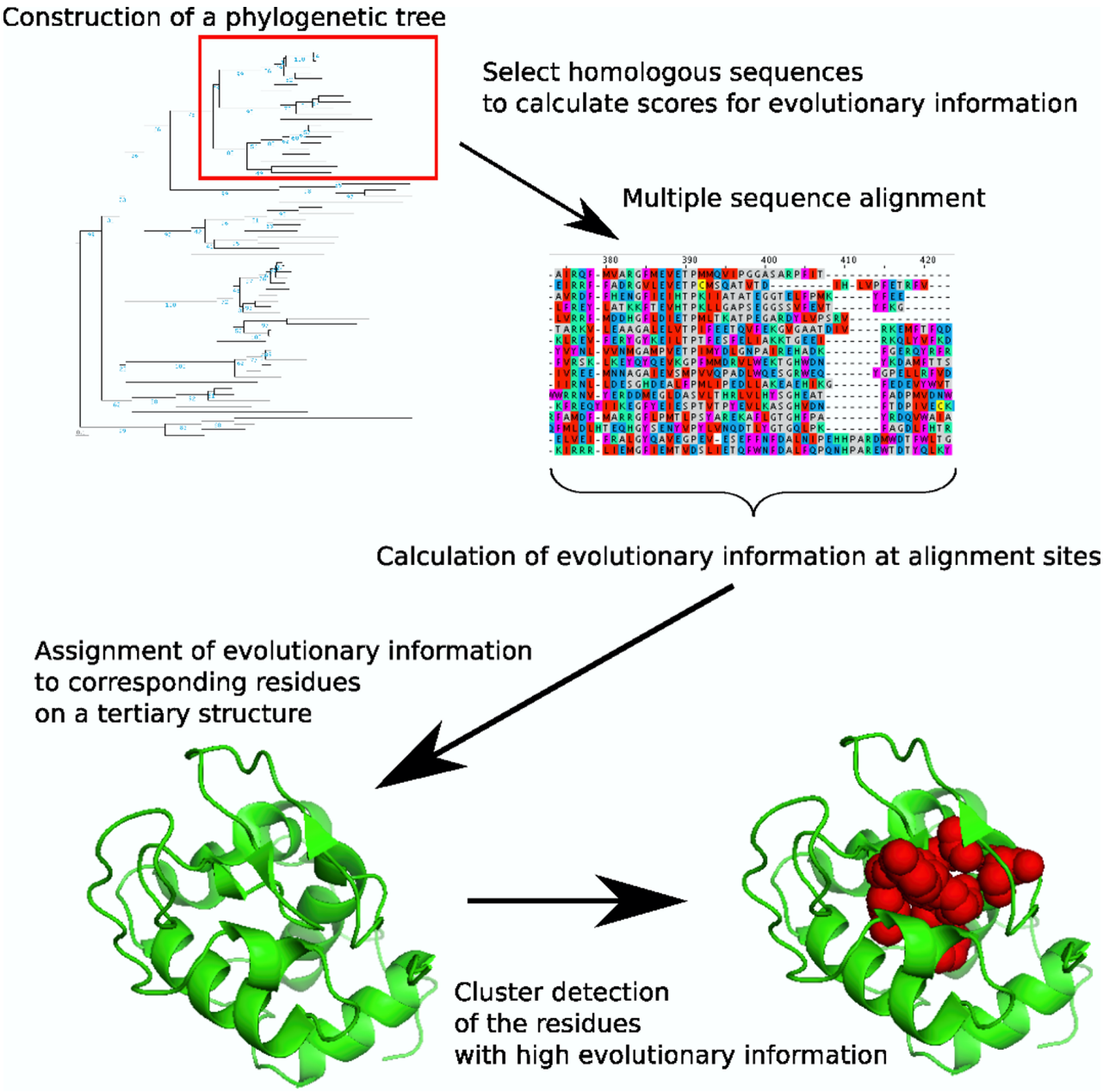

Figure 1. Procedure of the methods by integrating the structural and evolutionary information.

family are superimposed. Cheng *et al.* [16] also addressed a similar problem, and developed a method to predict the functional regions by distinguishing between functional constraints and structural constraints, but they adopted a different strategy to estimate the structural constraint. In order to obtain measurements of the structural constraints in a protein structure, they used Rosetta [17], which is a computational method to design a protein and calculate its free energy. They showed that combining these measures with sequence conservation improved the prediction of functional protein sites.

Zhang *et al.* [18] developed CUBE-DB, which provides calculated conservation and specialization scores for residues in paralogous proteins. The advantage of their database is that the functional specificity at a site is calculated by considering two models of evolution after divergence, "heterotachy" and "homotachy". The word heterotachy (for "different speed" in Greek) was applied by Lopez *et al.* [19] to refer to within-site rate variations throughout time in the field of molecular evolution. In contrast, homotachy (for "same speed" in Greek) refers to the state in which the evolutionary rate of a position is constant throughout time. Heterotachy was found among homologous sequences of distantly related organisms, often with different functions. In such cases, the functional constraints are likely to be distinct, which would explain the different distributions of variable sites. Zhang *et al.* [18] used heterotachy for referring to the evolutionary rate variations among homologous groups. A high score is calculated at a site where the residues are conserved in the reference group of orthologs, but they overlap poorly with the residue type choices in the paralogous groups (such positions are referred to as functional determinants). In contrast to the case of heterotachy, homotachy requires the conservation at a site within each paralogous group (referred to as functional discriminants). Residues with high scores are mapped on an evolutionarily related structure, if available, via Jmol [20], *etc.*, and are summarized as a table (html or downloadable xlsx format). CUBE-DB presently covers only human proteins belonging to multi-member families.