BioSystems 147 (2016) 67–77

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Profile comparison revealed deviation from structural constraint at the positively selected sites

Hiroyuki Oda[a,*], Motonori Ota[b], Hiroyuki Toh[c]

[a] Graduate School of Systems Life Sciences, Kyushu University, 744 Motooka Nishi-ku, Fukuoka 819-0395, Japan
[b] Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya City, Aichi 464-8601, Japan
[c] Department of Biomedical Chemistry, School of Science and Technology, Kwansei Gakuin University, 2-1 Gakuen, Sanda, Hyogo 669-1337, Japan

ARTICLE INFO

Article history:
Received 11 January 2015
Received in revised form 13 July 2016
Accepted 16 July 2016
Available online 18 July 2016

Keywords:
Positive selection
Molecular evolution
3D-profile
Position specific scoring matrices

ABSTRACT

The amino acid substitutions at a site are affected by mixture of various constraints. It is also known that the amino acid substitutions are accelerated at sites under positive selection. However, the relationship between the substitutions at positively selected sites and the constraints has not been thoroughly examined. The advances in computational biology have enabled us to divide the mixture of the constraints into the structural constraint and the remainings by using the amino acid sequences and the tertiary structures, which is expressed as the deviation of the mixture of constraints from the structural constraint. Here, two types of profiles, or matrices with the size of 20 x (site length), are compared. One of the profiles represents the mixture of constraints, and is generated from a multiple amino acid sequence alignment, whereas the other is designed to represent the structural constraints. We applied the profile comparison method to proteins under positive selection to examine the relationship between the positive selection and constraints. The results suggested that the constraint at a site under positive selection tends to be deviated from the structural constraint at the site.

1. Introduction

It is well known that the tertiary structures of proteins have been more conserved than the amino acid sequences during the course of molecular evolution (Chothia and Lesk, 1986; Russell and Barton, 1994). This observation suggested that the constraint to maintain protein folds is one of the major factors in the evolution of proteins, and hereafter they are referred to as "structural constraint". For example, the residues constituting the hydrophobic cores of globular proteins are subjected to strong structural constraints. On the other hand, it is also known that the constraints on the amino acid sites critical for the protein functions often deviate from the structural constraint (Elcock, 2001; Ota et al., 2003). Such examples can be found in the catalytic sites of enzymes. The amino acid residues constituting the catalytic sites are often polar or charged, even when these residues are located in the hydrophobic cavity. Despite the disadvantages from the structural viewpoint, such catalytic sites in the hydrophobic environment are invariant among the homologues (Ota et al., 2003). This means that the catalytic sites are subjected not only to structural constraint, but also to other types of constraints. In actual situation, the amino acid substitutions at each site in the primary structure of a protein are considered to have occurred by the influences of both the structural and the other types of constraints during the evolutionary process, but the relative intensities of these constraints are considered to differ from site to site.

Two groups have independently developed methods to evaluate the deviation of the constraint from the structural constraint at each site, by comparing two types of profiles. In the approach developed by Chelliah et al. (2004), both profiles are calculated from a multiple amino acid sequence alignment and take the same form, a matrix with the size of $20 \times L$, where L represents the alignment length. The (i, j) element of the matrix is the frequency of the i-th amino acid residue at the alignment site j. The elements of one of the profiles are directly calculated as the residue frequencies at each alignment site, whereas the elements of the other profile are calculated by taking the average of the substitution pattern of 20 amino acids over the residue type and the structural environment at the alignment site. The substitution pattern for a residue type and a structural environment is derived from the environment-specific substitution tables designed for local structural environments such as secondary structures and the accessibility of water molecules of the protein structure under consideration (Overington et al., 1992; Chelliah et al., 2004). A column of the former profile represents

* Corresponding author at: 2-3-24 Suidou, Bunkyou-ku, Tokyo 112-0005, Japan.
E-mail address: hiroyuki.oda1983@gmail.com (H. Oda).

http://dx.doi.org/10.1016/j.biosystems.2016.07.007
0303-2647/

the mixture of the constraints at the corresponding alignment site, whereas the corresponding column of the latter profile represents the structural constraints at the site. The deviation between the constraints at each alignment site is evaluated by the modified Kullback-Leibler divergence of the residue frequencies between the corresponding columns of the two profiles. On the other hand, Cheng et al. (2005) calculated two profiles as the position specific scoring matrices (PSSM), by the method used for PSI-BLAST (Altschul et al., 1997). One of the PSSMs is calculated from a multiple amino acid sequence alignment of homologous proteins. The other PSSM is calculated from an alignment of amino acid sequences, which are generated to fit a given tertiary structure by Rosetta Design (Kuhlman and Baker, 2000). A column of the former PSSM represents the mixture of the constraints at the alignment site, whereas a column of the latter PSSM represents only the structural constraints at the corresponding site. The deviation between the constraints at a site is evaluated by the Euclidian distance between the corresponding columns of the two PSSMs. Hereafter, this approach, which compares two types of profiles, is referred to as "profile comparison". Both groups applied their methods for the prediction of active sites (which includes catalytic sites and ligand binding sites). In their applications, the groups predicted that the sites with large deviations from the structural constraints would be active sites. However, the predictions with only the information about the deviations from the structural constraint did not show high performance. Chelliah et al. (2004) discussed the problem as follows: Enzymes often include sites under the other types of constraints, in addition to the active sites. For example, such sites include the residues involved in metal ion binding to maintain the local conformations of proteins, and those forming interfaces for protein-protein interactions. These sites are also subjected to the constraints related to the functions as well as structural constraints, although the reports about such sites are rare, comparing to those about the active sites. The situation is considered to be the cause of the false positives of their prediction (Chelliah et al., 2004). Therefore, the two groups included other measures, such as residue conservation and free energy, to improve the prediction performance. In this manuscript, we are interested in the application of profile comparison as a tool to evaluate the deviation from the structural constraint at each alignment site, rather than to predict active sites.

From an evolutionary viewpoint, residue conservation is observed at the sites under strong constraints, which have been maintained by purifying selection (also known as negative selection). That is, mutations at such sites would be eliminated rapidly if the mutated amino acid residues do not fit the constraints at the sites. As a consequence, the residues at such sites have been conserved. The active sites described above are often conserved based on the mechanism. Contrary to negative selection, positive selection is known as a driving force to accelerate amino acid substitutions, resulting in divergence at the sites. Quite a few examples of diverged sites under positive selection have been identified in various proteins. One well-known example is the variation at the antigenic determinant sites of pathogenic antigens (Agileta et al., 2009). Such sites are recognized by the host's antibodies. To escape from attacks by the host's immune system, the rate of amino acid substitutions at the antigenic determinant sites has been accelerated.

Roughly speaking, there are two approaches to detect positively selected sites. One of the approaches uses the ω ratio as a measure for the positive selection (Fitch et al., 1997; Nielsen and Yang, 1998; Suzuki and Gojobori, 1999; Yang et al., 2000; Kosakovsky Pond and Frost, 2005; Massingham and Goldman, 2005 Massingham and Goldman, 2005). The ω ratio is defined as the ratio of non-synonymous substitution rate to synonymous substitution rate. The synonymous substitution is a nucleotide substitution in a codon, which does not change the amino acid residue encoded by the codon. Therefore, the synonymous substitution is regarded as neutral at protein level. On the other hand, the non-synonymous substitution is a nucleotide substitution in a codon, which changes the encoded amino acid residue. The ω greater than 1.0 indicates that the amino acid substitution is accelerated comparing to the neutral change. Therefore, the site with $\omega > 1.0$ is considered to have evolved under positive selection. The other group considers that the changes in physicochemical character of amino acid residues at positively selected sites are larger than those at the remaining sites (Hughes et al., 1990; Rand et al., 2000; Zhang, 2000; Suzuki, 2007). They divide amino acid substitutions into two groups, "conservative" and "radical", based on the physicochemical properties of the amino acid residues. The former is the change between the residues with similar properties, whereas the latter is the change between the residues with quite different physicochemical properties, such as the substitution between neutral residues (e.g. leucine) and charged residues (e.g. arginine). Currently, the approach with the ω ratio is widely utilized for the detection of positively selected sites, and many sophisticated statistical methods have been developed for the approach (Yang and Nielsen, 2002; Wilson and McVean, 2006; Dutheil et al., 2012; Gharib and Robinson-Rechavi, 2013; Redelings, 2014; Zaheri et al., 2014; Angelis et al., 2014). As described above, the negative selection to generate residue conservation is associated with structural and/or functional constraints. On the other hand, the positive selection to generate residue divergence is associated with adaptation (e.g. Group II PLA_2, Lynch, 2007). It is unknown whether the amino acid substitution pattern at the positively selected sites is subject to either the structural constraint or other constraint than the structural one (e.g. functional constraints). In this manuscript, we have developed a new profile comparison method, and applied it to the characterization of the positively selected sites. At first, we examined whether our method can actually detect the deviation of constraint from the structural one, by applying the method to enzymes. We subsequently applied the method to proteins under positive selection. The positively selected sites were identified based on the ω ratio. We found that the constraints at the amino acid sites under positive selection tend to deviate from the structural constraints, as compared to the sites unrelated to the positive selection. If the amino acid substitutions according to the structural constraint are regarded as "conservative", and those that deviate from the structural constraints as "radical", then our study can be considered to connect the two approaches for the detection of positive selection, the approaches with the ω ratio (Fitch et al., 1997; Nielsen and Yang, 1998; Suzuki and Gojobori, 1999; Yang et al., 2000; Kosakovsky Pond and Frost, 2005; Massingham and Goldman, 2005) and those with the radical-conservative ratio (Hughes et al., 1990; Rand et al., 2000; Zhang, 2000; Suzuki, 2007).

2. Materials and methods

2.1. Dataset

2.1.1. Enzyme dataset

We used the enzymes registered in the Catalytic Site Atlas (CSA, http://www.ebi.ac.uk/thornton-srv/databases/CSA/, Porter et al., 2004), a database for the active sites of enzymes with available coordinates data. To construct reliable PSSMs, we selected the enzymes based on the number of available putative orthologous sequences. For the selection, we performed a BLAST search through the KEGG GENE DATABASE (http://www.genome.jp/kegg/genes.html, Kanehisa and Goto, 2000) with the amino acid sequence of each enzyme from CSA as a query. The putative orthologs were defined by the following four criteria: (1) The E-value is less than