Biochimica et Biophysica Acta 1824 (2012) 1484–1489

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

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

Structure-based rebuilding of coevolutionary information reveals functional modules in rhodopsin structure ☆

Keunwan Park, Dongsup Kim *

Department of Bio and Brain Engineering, KAIST, Daejeon, Republic of Korea

ARTICLE INFO

Article history:
Received 8 February 2012
Received in revised form 1 May 2012
Accepted 31 May 2012
Available online 8 June 2012

Keywords:
Correlated mutation analysis
Coevolution
Conservation
Rhodopsin
Functional site
Functional module

ABSTRACT

Correlated mutation analysis (CMA) has been used to investigate protein functional sites. However, CMA has suffered from low signal-to-noise ratio caused by meaningless phylogenetic signals or structural constraints. We present a new method, Structure-based Correlated Mutation Analysis (SCMA), which encodes coevolution scores into a protein structure network. A path-based network model is adapted to describe information transfer between residues, and the statistical significance is estimated by network shuffling. This model intrinsically assumes that residues in physical contact have a more reliable coevolution score than distant residues, and that coevolution in distant residues likely arises from a series of contacting and coevolving residues. In addition, coevolutionary coupling is statistically controlled to remove the structural effects. When applied to the rhodopsin structure, the SCMA method identified a much higher percentage of functional residues than the typical coevolution score (61% vs. 22%). In addition, statistically significant residues are used to construct the coevolved residue–residue subnetwork. The network has one highly connected node (retinal bound Lys296), indicating that Lys296 can induce and regulate most other coevolved residues in a variety of locations. The coevolved network consists of a few modular clusters which have distinct functional roles. This article is part of a Special Issue entitled: Computational Methods for Protein Interaction and Structural Prediction.

1. Introduction

The most widespread approach for finding protein functional sites is to examine conservational positions within the homologous sequences generated by multiple sequence alignment (MSA) [1–4]. Mutations on conserved residues have been thought to cause deleterious effects, which may explain the conservational phenomenon that occurs under long evolutionary pressure. Indeed, much previous research has supported the idea [2,5]. However, functional residues are not always conserved; they often mutate for a variety of reasons, such as altering the functional specificity [6–8], maintaining structural stability [9], and inducing allosteric signals (e.g. conformational epitasis) [10]. Thus, analyzing variable positions can provide another opportunity to find functional sites uncovered by conservational information. Correlated mutation analysis (CMA) deals with variable positions in MSA and identifies coevolved patterns between MSA positions. Until now, many approaches have connected coevolutionary information to various biological functions [6–9,11–14]. However, a weakness of CMA is that coevolved patterns contain heterogeneous signals, such as structural complementarities [12], phylogenetic noises [11], and functional interactions [6,8,12]. Thus, a method is required for extracting only the functional information from these various signals.

In this study, we developed a new method called structure-based correlated mutation analysis (SCMA), which encodes typical coevolutionary information into a protein structure network. A path-based network model was adapted to model information transfer between residues, and the statistical significance of the transferred information was estimated by empirical cumulative density function (ECDF) derived from network shuffling. The SCMA method intrinsically assumes that the coevolution score of residues in physical contact is more reliable than that of distant ones, and that distant coevolution likely arises from a series of contacting and coevolving residues.

The SCMA method was applied to analyze the rhodopsin structure (PDB ID: 2J4Y) because of its biological importance, well-characterized structure, and sufficient sequence information. The results showed that the SCMA method identified considerably more functional residues than the typical coevolution score (61% vs. 22%). In addition, the statistically significant residue pairs (p-value < 0.001) were used to construct a coevolved residue–residue subnetwork. Specifically, we rebuilt the coevolutionary relationship by modeling long-range residue–residue interactions, thus creating the subnetwork containing the distant functional residues. Interestingly, this network had one highly connected node—retinal bound Lys296—indicating that Lys296 could

☆ This article is part of a Special Issue entitled: Computational Methods for Protein Interaction and Structural Prediction.

* Corresponding author at: Department of Bio and Brain Engineering, KAIST, 373-1, Guseong-dong, Yuseong-gu, Daejeon, 305-701, Republic of Korea. Tel.: +82 42 350 4317; fax: +82 42 350 4310.
E-mail addresses: paintzzz@kaist.ac.kr (K. Park), kds@kaist.ac.kr (D. Kim).

1570-9639/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbapap.2012.05.015

induce and regulate most other coevolved residues in a variety of locations, including the cytoplasmic domain, far from the Lys296. Moreover, the coevolved network consisted of a few modular clusters (i.e. protein sectors [15]) which had distinct functional roles in the rhodopsin structure. Finally, little overlap occurred between the coevolutionary and conservational residues, representing different coverage between the two.

2. Results and discussion

2.1. Rhodopsin structure of G protein-coupled receptor (GPCR) family

Rhodopsin, a GPCR family member, is the mammalian dim light photoreceptor in the outer segments of rods, wherein the ligand (11-cis retinal) is covalently attached to Lys296 via Schiff base linkage. The rhodopsin structure includes three types of domain: intradiscal (corresponding to the extracellular domain of related GPCRs), membrane-embedded, and cytoplasmic. The intradiscal domain is mainly composed of loops connecting the seven transmembrane helices (H1-H7) of the membrane-embedded domains [16]. The cytoplasmic domain also consists of many loops, as well as a short α-helix (H8) lying nearly perpendicular to H7, which is likely involved in G-protein activation. Upon light-activation, 11-*cis*-retinal isomerizes to all-*trans*-retinal, resulting in a series of intermediate conformational states. This visual signal transduction around Lys296 induces subsequent interactions in the distant cytoplasmic domain, which in turn lead to a variety of downstream signaling cascades [17].

2.2. Rebuilding coevolutionary information based on residue–residue contacts to detect long-range functional coupling

In this section, the SCMA method is briefly explained. See the Method section for the detailed calculation procedure with a simple example. The input of the SCMA method is the MSA of the protein structure. The sequence information of the structure is used to construct the MSA using iterative psi-blast search with nr database. The raw coevolution score is also calculated from the MSA and used to construct the network model. That is, the protein structure network is generated with nodes (residues) that are connected at the edges, where the mutual distance is less than 5 Å. Next, the edge weight is assigned as the raw coevolution score between the residues. Note that only the relative coevolution score is important for the subsequent procedure, rather than the absolute value. After that, residue–residue coevolution is rebuilt by modeling information transfer between residues in the protein structure network (Fig. 1).

Conceptually, the SCMA method gives a high score to a residue pair that features many alternative paths consisting of a series of coevolving and contacting residues. Therefore, the method intrinsically assumes that the coevolution score of residues in physical contact is more reliable than that of distant ones, and that the functional coevolution between distant residues likely arises from a physically-connecting residue network connecting the two sites.

In addition, the residue coupling (i.e. information transfer in the network model) is statistically controlled to remove the structural (noisy) effects on coevolution. Within a fixed network architecture, shuffling the edge weights allows us to estimate how statistically

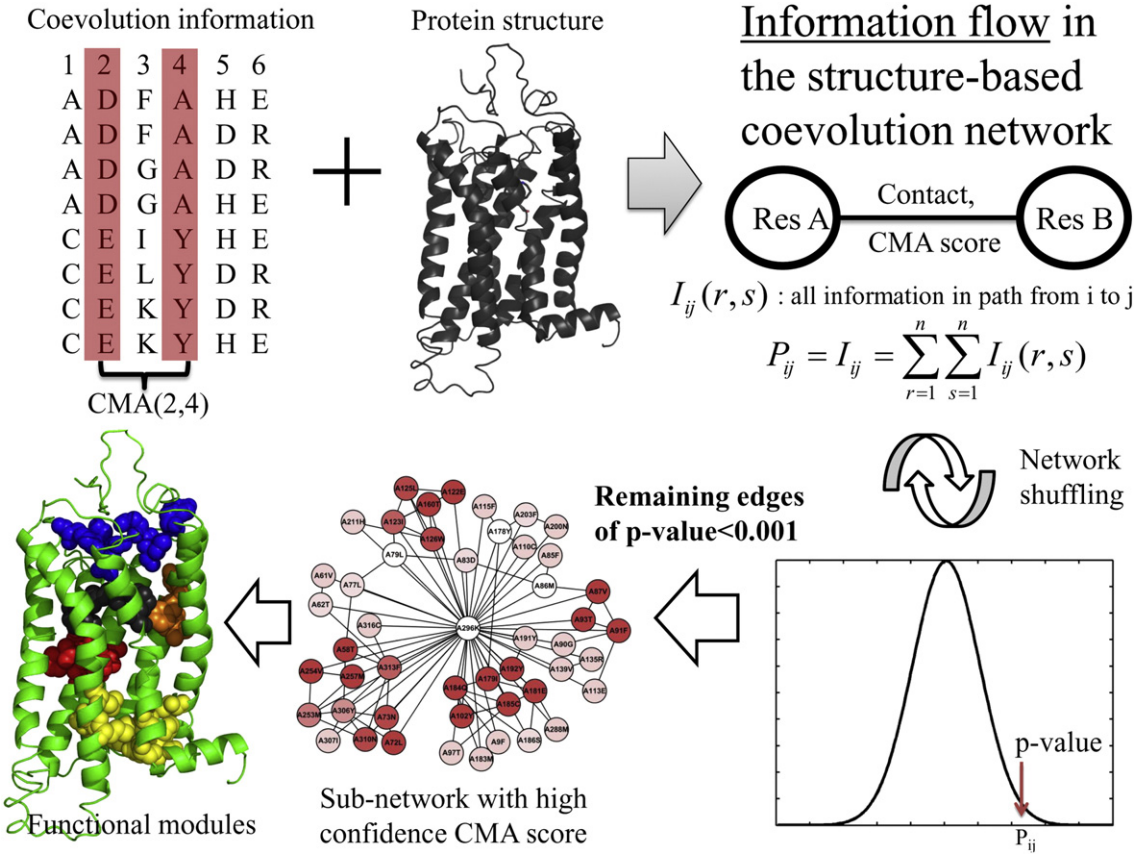

Fig. 1. The overall procedure of structure-based correlated mutation analysis (SCMA) method is shown. The coevolutionary information and protein structure information are combined to construct a coevolutionary contact network. In this network, a node is defined as a residue, and an edge weight as a coevolution score between contact residues (<5 Å). Based on the network, transferred information between residues is calculated by considering the reciprocal of the variances for all possible paths. The statistical significance of the corresponding transferred information is estimated by network shuffling and then used to construct the functionally coevolved subnetwork.