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)

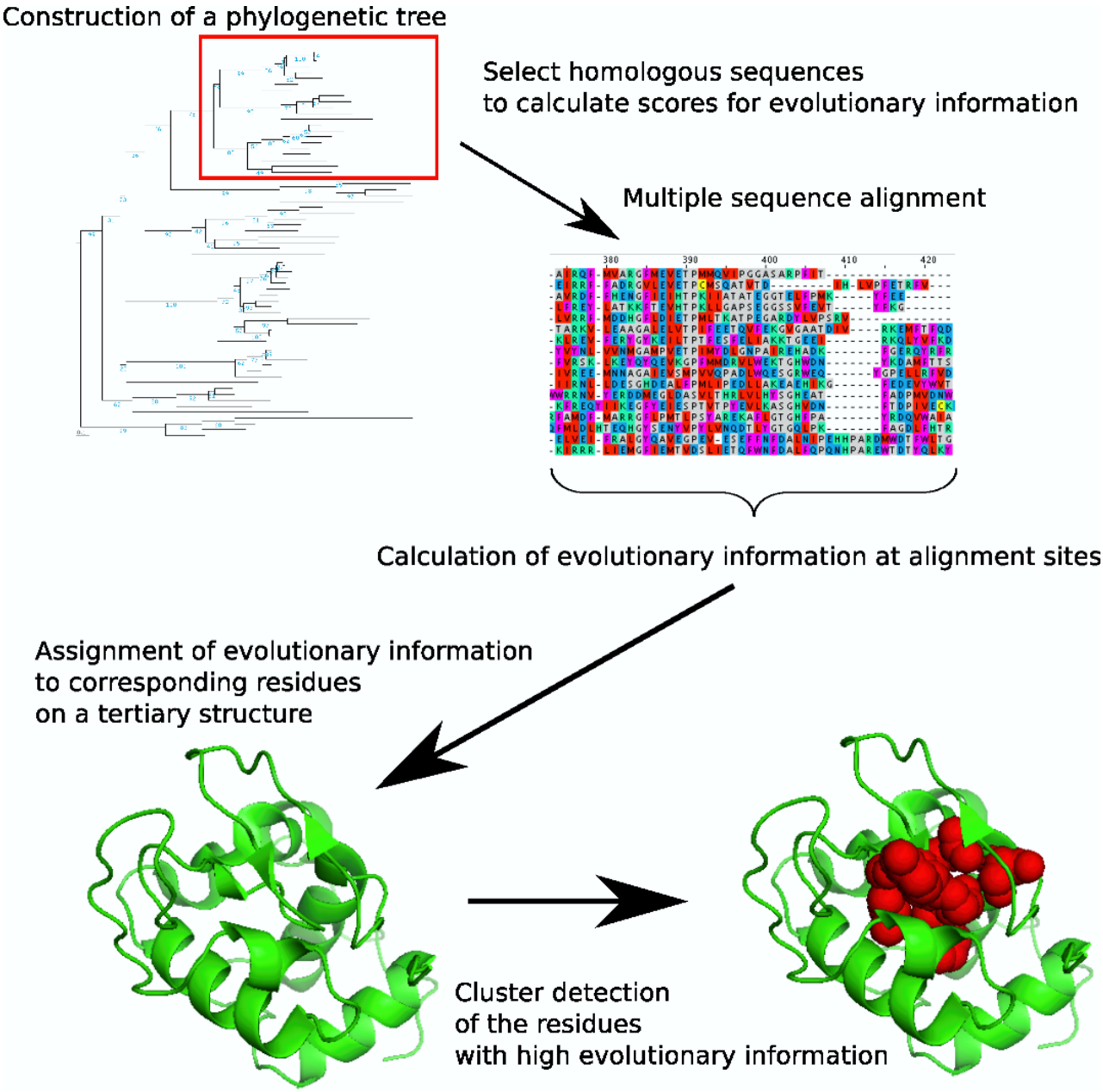

Figure 1. Procedure of the methods by integrating the structural and evolutionary information.

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.

Table 1. The computational methods to predict functional regions by using evolutionary and structural information, discussed in this review, which are available through the internet. Names (abbreviation) and classification of each service (Server or Database), URLs, and their features are described in each column.

Name (Server/Database)	URL	Description
ET (Server)	http://mammoth.bcm.tmc.edu/ETserver.html	**Pioneering work** Last updated: 2011
Crescendo (Server)	http://mordred.bioc.cam.ac.uk/~crescendo/crescendo.php	**Discrimination between structural and functional constraints** Structural and functional constraints are discriminated by the application of ESST. Last updated: 2006
CUBE-DB (Database)	http://epsf.bmad.bii.a-star.edu.sg/cube/db/html/home.html	**Integration of conservation & other scores** Pre-evaluated conservation and specialization scores for residues in paralogous proteins are provided as a table. Last updated: 2012
JET (Server)	http://www.lgm.upmc.fr/JET/JET.html	**Integration of conservation & other scores** Integrated scores of residue conservation and physicochemical properties are used for the prediction. Last updated: 2009
SitesIdentify (Server)	http://www.manchester.ac.uk/bioinformatics/sitesidentify/	**Conservation & other scores** Conservation scores and geometry-based cleft identification are used for the prediction. This server may no longer be updated.
Direct Coupling Analysis (DCA)	http://dca.upmc.fr/DCA/DCA.html	**Coevolutionary relationship between two sites** Discrimination between directly and indirectly correlated residues was achieved by direct-coupling analysis (DCA). Last updated: 2012
MISTIC (Server)	http://mistic.leloir.org.ar/index.php	**Visualization of coevolving sites** Connectivity among coevolving sites is visualized by a circular representation of the MI network. Last updated: 2013
ETA (Server)	http://mammoth.bcm.tmc.edu/eta/	**Toward function prediction** A 3D template composed of ET residues is used for function prediction. Last updated: 2013
FREPS (Server)	http://freps.cbrc.jp	**Appropriate sequence selection** The MSA with the maximum DSPAC is adopted for the prediction. Last updated: 2012
FunShift	http://funshift.sbc.su.se	**Functional shift analysis** FunShift performs functional shift (divergence) analysis between the subfamilies of a protein domain family. The present release uses Protein Domain families in Pfam (Version 12.0). Last updated: 2004
EvoDesign (Server)	http://zhanglab.ccmb.med.umich.edu/EvoDesign/	**Protein design** Designing ideal protein sequences for given scaffolds by evaluating the foldability and goodness of the designs. Last updated: 2013

Integration of conservation and other scores

It is difficult to predict all of the types of functional regions by a single score, because each functional region has its own physico-chemical properties. For example, protein interfaces are not simply discriminated from non-interface surfaces by the patches of conserved residues [21]. The most conserved patches of residues overlap in only 37.5% of the actual protein interface, although the properties of the interface differ from those of the rest of the protein. Considering other types of scores seems to be essential, to improve the prediction accuracy. Engelen *et al.* [22] addressed this problem by integrating the conservation information and the specific physicochemical properties of the residues. They developed the Joint Evolutionary Trees (JET) method [22], to detect protein interfaces, the core residues involved in the folding process, and the residues susceptible to site-directed mutagenesis and relevant to molecular recognition. The performance of JET is better than those of the other state-of-the-art methods.

Teppa *et al.* [23] compared the abilities of several methods (real-valued ET [24], cumulative mutual information (MI) [25], proximity MI [25], evolutionary trace integer value [2], and the methods designed for the identification of SDPs (SDPfox and XDET) [26]) to identify catalytic residues in enzymes, in order to investigate the extents to which the predictive powers of the different methods overlap. The results revealed that the methods can be divided in three groups, with limited mutual overlaps. These groups consist of the methods in which the predictive signal is strongly correlated to the sequence conservation, those in which the predictive signal is derived from MI, and those developed for the prediction of specificity-determining positions. Interestingly, the combined scores of the first and second groups (sequence conservation group and MI group) achieved the highest performance. These observations revealed that the sequence conservation and the MI scores are considered to be distinct signals encoded on the MSA, and produce a complementary effect. Therefore, their results demonstrated the possibility of detecting catalytic residues more accurately, by integrating structural and higher-order sequence evolutionary information. Thus, the integration of the conservation score with other types of scores represents a trend toward improving evolutionary information methods. In addition to the methods examined in Teppa's work [2,24–26], Bray *et al.* [27] developed a functional site prediction tool, SitesIdentify, which is based on combining sequence conservation information with geometry-based cleft identification. This method functioned quite favorably in comparison to other methods, in the active site predictions for 237 non-redundant enzymes. As of 1st November, 2013, the SitesIdentify server is not working at the URL described in the original paper [27].

Coevolutionary relationship between two sites

The evolutionary scores calculated in an MSA used for functional region prediction are roughly divided into two types: the scores at a site and those between sites. The former type is the conservation/variation of amino acids at a site, while the latter one is the score of the degree of coevolutionary relationship between two or more sites. In our opinion, the performance comparisons by Teppa *et al.* [23], described in the previous section, only focused on the former point: the conservation/variation of amino acids at a site. Several methods, such as correlated mutation [28,29], MI [30–36], and covariation [37], have been developed to estimate the degree of the coevolutionary relationship between two sites among the sequences in an MSA. The methods in the former group are based on conserved residues in MSAs, but the methods in the latter group that detect coevolution between two sites are based on variable sites. Furthermore, it should be noted that invariable sites do not contain any information in the coevolutionary score. A high coevolutionary score is considered to indicate the spatial proximity or functional connectivity between the sites, even when the sites are not close in the primary structure of a protein. Therefore, these methods have been applied to detect not only intramolecular interactions [28–38] but also intermolecular interactions [39], regardless of direct or indirect interactions. Aguilar *et al.* [40] investigated how coevolution information can be used to improve the prediction methods for functional residues. They found that the clusters of co-evolving sites related to the catalytic sites of an enzyme have distinguishable topological properties in the residue-residue interaction network, and also observed that these clusters usually evolve independently. Interestingly, they suggested that the clustering of coevolving residues could be related to a fail-safe mechanism, which causes no harm or minimizes harm to other parts in a protein structure, in the case of a functional loss at a site. Kowarsch *et al.* [41] performed comprehensive analyses of point mutations causing human diseases, with respect to the correlated mutations. They showed that 1) the correlated sites are significantly more likely to be disease-associated than expected, 2) these signals cannot be explained by the conservation patterns at each site, and 3) many correlations are not related to physical contacts between sites. However, Halperin *et al.* [42] highlighted the limitation of correlated mutation analyses, which might also be true for other coevolutionary relationship-based approaches. They showed that several correlated mutation methods achieve practical accuracy for intramolecular interaction prediction on their dataset, but the accuracy declines for intermolecular interaction prediction. Overall, they insisted that the examined methods are not suitable for large-scale intermolecular contact predictions. In other words, the current methods can only achieve practical accuracy for a handful of families. Therefore, the potential for the application of coevolutionary information to functional region prediction remains debatable. In addition, these methods have an important shortcoming that is considered to affect their predictive accuracies. One important problem stems from the fact that correlation in amino acid substitution may arise from direct as well as indirect interactions [43].

The availability of many protein sequences enables the use of various statistical approaches to address this problem. Recently, discrimination between directly and indirectly correlated residues was achieved by the direct-coupling analysis (DCA) by Weigt *et al.* [43]. DCA combines covariance analysis with global inference analysis, adopted from use in statistical physics. A message-passing algorithm was used to implement DCA (mpDCA), but it was rather costly computationally because it is based on a slowly converging iterative scheme. Hence, the same group applied an algorithm based on the mean-field approximation of DCA (mfDCA), which is 103 to 104 times faster than mpDCA, and thus can be used to analyze many long protein sequences rapidly [44]. In addition to the DCA-related methods, several groups addressed the discrimination between direct and indirect correlations [45–50]. These methods were primarily applied to identify constraints to fold a protein tertiary structure [44–49]. However, Weigt *et al.* [43] applied their method to identify constraints to maintain a protein-protein interaction [43,51].

In addition, one of the recent advances in coevolution-based approaches is the development of MISTIC [52], by which a user can visualize the connectivity among coevolving sites as a circular representation of an MI network and their MIs interactively. Even when the initial result is too complicated to understand, other scores (cumulative MI [25], conservation, proximity MI [25] *etc.*) can be considered simultaneously at both the nodes and edges, to highlight the information encoded within an MSA. Such a visualization tool clarifies the intricate evolutionary connections among sites.

Toward function prediction

Almost all annotations assigned to protein sequences rely primarily on the computational identification of similarity between the protein sequences with unknown and known functions, which are identified by BLAST [53] and other programs. Annotations are often misleading when the sequence similarities between the queries and the retrieved sequences are low. In these cases, the global structural similarity between a protein of unknown function and one with a known function is utilized [54–62]. Even when this approach fails, further functional information might be obtained by using local structural similarities [63,64]. For example, surface patches or clefts [65–73], or tertiary templates of small numbers of functional residues [65,74–81] are used as 3D templates to infer protein functions, through the identification of the corresponding key functional residues and their geometries on other structures. Such 3D templates may identify functional analogs without detectable homology that convergently perform the same function. Various characteristics of local structures have been used as queries to identify local structural similarity in distantly related- or non-homologous structures. These functional region predictions by local structural matches are often complementary to methods by global structural or sequence-based matches, when global structural or sequence-based methods do not provide detailed information about a protein of unknown function. In our opinion, however, evolutionary information has not been fully exploited for the detection of functional similarity among non-homologous structures.

Recently, Kristensen *et al.* [82] applied evolutionary information to develop the Evolutionary Trace Annotation (ETA) pipeline, which in principle can be applied to detect functional similarity among non-homologous structures. The basic idea of the ETA-based function annotation [82] is described as follows. A schematic image of the method is available at http://mammoth.bcm.tmc.edu/eta/manual.html. At first, a few key functional residues are clustered into 3D templates, and local structures similar to them are searched for in other protein structures. Secondly, when the 3D template is matched in the structure of a protein, the function of the structure in the found protein is transferred to the query structure. In order to increase the sensitivity of the functional annotation [83], the proteins identified by ETA are linked together into a network of ETA similarities; then, starting from proteins with known functions, competing functional labels diffuse link-by-link over the entire network. A likelihood *z*-score for every function is assigned to every node. The function corresponding to the most significant score is adopted at each node, as its annotation, for example, by referring to the Enzyme Commission (EC) numbers of the retrieved structures. In high throughput controls, this competitive diffusion process recovered enzyme activity annotations with 99% and 97% accuracies at half-coverage for the third and fourth levels of the EC number, respectively, although currently these predictions have only been evaluated for homologs. These accuracies corresponded to false positive rates 4-fold lower than that of the nearest-neighbor method and 5-fold lower than that of the sequence-based annotations.

Automatic sequence selection to evaluate evolutionary information

The selection of homologous sequences is a critical step in the prediction of functional regions by using the conservation score, because conserved residues are identified through comparisons of homologous sequences [84]. The same is true for the other evolutionary information-based methods described above. We empirically know that a certain degree of sequence divergence in the set of homologous sequences is essential for the identification of conserved residues. However, the selection of an appropriate homologous sequence to calculate residue conservation has not been sufficiently addressed. Aloy *et al.* [8] developed an automatic method to predict the functional regions of a protein, by detecting conserved residue clusters on the tertiary structure. If no cluster is identified, then the MSA is reconstructed by removing the distant homologues to the prediction target, according to the evolutionary relationships suggested by a phylogenetic tree. The process is iterated until at least one cluster is identified. In other words, the iteration process is forcibly terminated, even if more appropriate conditions are present within the untested sequence space. Mihalek *et al.* [85,86] applied a residue clustering measure, which was originally developed as a formula for the identification of conserved residue clusters, to indicate the appropriateness of a set of sequences for functional region predictions [85,86]. The measure quantifies the degree of clustering of the evolutionarily important residues in the tertiary structure of a protein, and attaches greater importance to the clustering of the residues that are far from each other on the primary structure. The sequence set selected by their measure performed better in their functional region prediction by the real valued ET-based method [87]. They showed that the performance of their method for protein-protein interaction interfaces was lower than that for active sites. Recently, we addressed a similar problem by a different approach, and developed the Functional REgion Prediction by using Spatial statistics (FREPS) method [84]. FREPS implements an index, DSPAC (the Degree of Spatial AutoCorrelation), to measure the appropriateness of a set of homologous sequences [84]. Structure and sequence information are integrated by spatial statistics within the index, which represents the degree of conserved residue clustering on the tertiary structure of the protein. The functional region prediction performance, using the set of sequences selected by DSPAC, was better than that obtained using the set selected under the fixed percent sequence identity-conditions. In addition, DSPAC successfully distinguished the sequence set including only C-type lysozyme from that including both C-type lysozyme and its non-enzyme homologue, α-lactalbumin. Similar to the residue clustering measure [85,86], however, the performance of DSPAC for protein-protein interaction interfaces was lower than that for active sites, although the details have not been published yet. In order to assess various types of functional regions, DSPAC [84] and residue clustering measures [85,86] should be improved.

Future directions

We would like to conclude by describing three possible extensions of the methods using structure and evolutionary information. The first is the extension to identify the functional differences of closely related proteins. In considerable numbers of protein families, a subfamily develops a new function, changes substrate specificity, or loses an original function. These family members can be categorized into several subfamilies, for investigations of their functions. Several databases and methods have been developed. For example, the FunShift database [88] is a collection of such functionally changed subfamilies (function shift) in a Pfam [89] protein family, identified by using Conservation Shifting Sites and Rate Shifting Sites. It is useful for protein design and mutagenesis studies, although FunShift has not been updated since 2004. PANTHER [90] also provides similar information curated by experts. Recently, Lee *et al.* developed GeMMA [91], which automatically classifies families and superfamilies into functional subfamilies, and is comparable to the established method SCI-PHY [92]. The common feature of these methods and databases is that they do not consider structural

information, which might limit their predictive accuracies. As described above and in our previous work [84], DSPAC was able to distinguish the sequence set including only the proteins with identical functional structures, from that including both the proteins with identical functions and those with different functions. Theoretically, the same is expected to be true for the residue clustering measure [85,86]. In addition, the ETA-based search [82] could be extended to identify the functional differences by considering the local structural differences between a 3D template and a matched one. These measures might improve the predictive accuracy to identify the functional differences of closely related proteins.

The second point is the extension to the template selection in homology modeling, by the application of evolutionary information. See details in our previous work [84]. Homology modeling is frequently utilized for a sequence without a solved crystal structure. Thus, the main purpose of the modeling is to investigate the molecular function. It would be better to use the structure of a protein with a function considered to be the same as or similar to the sequence under consideration as a template, although a model based even on a template with different functions can provide important information. However, the template structures retrieved by fold recognition programs do not always have the identical or similar function to those of the target sequence, because such programs do not directly evaluate the functional similarity between the target and a retrieved structure. Here, we consider the inverse problem. Suppose that we have a structure. The problem is to determine which homologous sequences can be modeled, using the given structure as the template. Fold recognition programs may not provide an answer to the problem, since the functional similarity to the structure-known protein is not considered in these programs. Considering the functional differences by an ETA-based strategy [82], residue clustering measures [85,86] or DSPAC [84] can be used to solve the problem, since these can be extended to identify the set of sequences that shares the same or similar biochemical functions.

The third point is the extension to protein design [93], which is becoming a popular research subject. One of the main purposes of protein design is to optimize the physicochemical characteristics of a designed structure. Most current protein design methods rely on physics-based force fields to search for low free-energy states. Recently, Mitra *et al.* [94] developed a method, EvoDesign, to design ideal protein sequences for given scaffolds. At first, EvoDesign collects a set of proteins with similar folds from the PDB data, by a structural alignment with a query structure. Then, an evolutionary profile is constructed from the MSA of the retrieved structures. This profile is used for a conformational search in sequence space, where the physicochemical packing of the side-chain and backbone atoms is accommodated by neural-network-based solvation, torsion angle and secondary structure predictions. However, this step does not consider the functional similarity with the scaffold, but mainly focuses on the foldability and goodness of the designs. Therefore, the retrieved structures might not always have the identical or similar function to that of the original scaffold. Evaluating the functional similarity by evolutionary information-based approaches would contribute to progress in protein design.

Acknowledgements

WN was supported by a Grant-in-Aids for Young Scientists B (25870764) from the Japan Society for the Promotion of Science (JSPS), and by a Grant from the Research Institute for Science and Technology of Tokyo Denki University (Q13L-03).

Author contributions

WN designed the concept of the mini review, and wrote the paper. AS and HO collected information about several web services, used them to analyze data, and discussed their advantages and disadvantages with WN.

Citation

Nemoto W, Saito A, Oikawa H (2013) Recent advances in functional region prediction by using structural and evolutionary information – Remaining problems and future extensions. Computational and Structural Biotechnology Journal. 8 (11): e201308007. doi: http://dx.doi.org/10.5936/csbj.201308007

References

1. Watson JD, Laskowski R a, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15: 275–284.
2. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257: 342–358.
3. Landgraf R, Xenarios I, Eisenberg D (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 307: 1487–1502.
4. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, et al. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 316: 139–154.
5. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, et al. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19: 163–164.
6. Chelliah V, Chen L, Blundell TL, Lovell SC (2004) Distinguishing structural and functional restraints in evolution in order to identify interaction sites. J Mol Biol 342: 1487–1504.
7. Del Sol A, Pazos F, Valencia A (2003) Automatic methods for predicting functionally important residues. J Mol Biol 326: 1289–1302.
8. Aloy P, Querol E, Aviles FX, Sternberg MJ (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 311: 395–408.
9. Innis CA, Anand AP, Sowdhamini R (2004) Prediction of functional sites in proteins using conserved functional group analysis. J Mol Biol 337: 1053–1068.
10. Mihalek I, Res I, Lichtarge O (2006) Evolutionary trace report_maker: a new type of service for comparative analysis of proteins. Bioinformatics 22: 1656–1657.
11. Joachimiak MP, Cohen FE (2002) JEvTrace: refinement and variations of the evolutionary trace in JAVA. Genome Biol 3: research0077.1–research0077.12.
12. Innis CA, Shi J, Blundell TL (2000) Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. Protein Eng 13: 839–847.
13. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38: W529–33.
14. Valdar WSJ (2002) Scoring residue conservation. Proteins 48: 227–241.

15. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci 1: 216–226.
16. Cheng G, Qian B, Samudrala R, Baker D (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res 33: 5861–5867.
17. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97: 10383–10388.
18. Zhang ZH, Bharatham K, Chee SMQ, Mihalek I (2012) Cube-DB: detection of functional divergence in human protein families. Nucleic Acids Res 40: D490–4.
19. Lopez P, Casane D, Philippe H (2002) Heterotachy, an important process of protein evolution. Mol Biol Evol 19: 1–7.
20. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/ (n.d.).
21. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13: 190–202.
22. Engelen S, Trojan L a, Sacquin-Mora S, Lavery R, Carbone A (2009) Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. PLoS Comput Biol 5: e1000267.
23. Teppa E, Wilkins AD, Nielsen M, Buslje CM (2012) Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction. BMC Bioinformatics 13: 235.
24. Mihalek I, Res I, Yao H, Lichtarge O (2003) Combining inference from evolution and geometric probability in protein structure evaluation. J Mol Biol 331: 263–279.
25. Marino Buslje C, Teppa E, Di Doménico T, Delfino JM, Nielsen M (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. PLoS Comput Biol 6: e1000978.
26. Mazin P V, Gelfand MS, Mironov AA, Rakhmaninova AB, Rubinov AR, et al. (2010) An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. Algorithms Mol Biol AMB 5: 29.
27. Bray T, Chan P, Bougouffa S, Greaves R, Doig AJ, et al. (2009) SitesIdentify: a protein functional site prediction tool. BMC Bioinformatics 10: 379.
28. Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18: 309–317.
29. Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7: 349–358.
30. Gloor GB, Martin LC, Wahl LM, Dunn SD (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 44: 7156–7165.
31. Martin LC, Gloor GB, Dunn SD, Wahl LM (2005) Using information theory to search for co-evolving residues in proteins. Bioinformatics 21: 4116–4124.
32. Tillier ERM, Lui TWH (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics 19: 750–755.
33. Yip KY, Patel P, Kim PM, Engelman DM, McDermott D, et al. (2008) An integrated system for studying residue coevolution in proteins. Bioinformatics 24: 290–292.
34. Gouveia-Oliveira R, Roque FS, Wernersson R, Sicheritz-Ponten T, Sackett PW, et al. (2009) InterMap3D: predicting and visualizing co-evolving protein residues. Bioinformatics 25: 1963–1965.
35. Fares MA, McNally D (2006) CAPS: coevolution analysis using protein sequences. Bioinformatics 22: 2821–2822.
36. Kozma D, Simon I, Tusnády GE (2012) CMWeb: an interactive online tool for analysing residue-residue contacts and contact prediction methods. Nucleic Acids Res 40: W329–33.
37. Larson SM, Di Nardo AA, Davidson AR (2000) Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J Mol Biol 303: 433–46.
38. Nemoto W, Imai T, Takahashi T, Kikuchi T, Fujita N (2004) Detection of pairwise residue proximity by covariation analysis for 3D-structure prediction of G-protein-coupled receptors. Protein J 23: 427–435.
39. Filizola M, Olmea O, Weinstein H (2002) Prediction of heterodimerization interfaces of G-protein coupled receptors with a new subtractive correlated mutation method. Protein Eng 15: 881–885.
40. Aguilar D, Oliva B, Marino Buslje C (2012) Mapping the mutual information network of enzymatic families in the protein structure to unveil functional features. PLoS One 7: e41430.
41. Kowarsch A, Fuchs A, Frishman D, Pagel P (2010) Correlated mutations: a hallmark of phenotypic amino acid substitutions. PLoS Comput Biol 6.
42. Halperin I, Wolfson H, Nussinov R (2006) Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins 63: 832–845.
43. Weigt M, White R a, Szurmant H, Hoch J a, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A 106: 67–72.
44. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A 108: E1293–301.
45. Marks DS, Colwell LJ, Sheridan R, Hopf T a, Pagnani A, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6: e28766.
46. Taylor WR, Sadowski MI (2011) Structural constraints on the covariance matrix derived from multiple aligned protein sequences. PLoS One 6: e28265.
47. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys 87: 012707.
48. Jones DT, Buchan DW a, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28: 184–190.
49. Cocco S, Monasson R, Weigt M (2013) From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction. PLoS Comput Biol 9: e1003176.
50. Miyazawa S (2013) Prediction of contact residue pairs based on co-substitution between sites in protein structures. PLoS One 8: e54252.
51. Lunt B, Szurmant H, Procaccini A, Hoch JA, Hwa T, et al. (2010) Inference of direct residue contacts in two-component signaling. Methods Enzymol 471: 17–41.
52. Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C (2013) MISTIC: mutual information server to infer coevolution. Nucleic Acids Res 41: W8–W14.
53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

54. Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233: 123–138.
55. Madej T, Gibrat JF, Bryant SH (1995) Threading a database of protein cores. Proteins 23: 356–369.
56. Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr Sect D Biol Crystallogr 60: 2256–2268.
57. Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, et al. (2003) Recognizing the fold of a protein structure. Bioinformatics 19: 1748–1759.
58. Ausiello G, Zanzoni A, Peluso D, Via A, Helmer-Citterich M (2005) pdbFun: mass selection and fast comparison of annotated PDB residues. Nucleic Acids Res 33: W133–W137.
59. Gilbert D, Westhead D, Nagano N, Thornton J (1999) Motif-based searching in TOPS protein topology databases. Bioinformatics 15: 317–326.
60. Jambon M, Imberty A, Deléage G, Geourjon C (2003) A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 52: 137–145.
61. Jambon M, Andrieu O, Combet C, Deléage G, Delfaud F, et al. (2005) The SuMo server: 3D search for protein functional sites. Bioinformatics 21: 3929–3930.
62. Lisewski AM, Lichtarge O (2006) Rapid detection of similarity in protein structure and function through contact metric distances. Nucleic Acids Res 34: e152.
63. Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, et al. (2002) PROSITE: A documented database using patterns and profiles as motif descriptors. Br Bioinform 3: 265–274.
64. Nevill-Manning CG, Wu TD, Brutlag DL (1998) Highly specific protein sequence motifs for genome analysis. Proc Natl Acad Sci U S A 95: 5865–5871.
65. Barker JA, Thornton JM (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics 19: 1644–1649.
66. Kleywegt GJ, Jones TA (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. Acta Crystallogr Sect D Biol Crystallogr 50: 178–185.
67. Shulman-Peleg A, Nussinov R, Wolfson HJ (2004) Recognition of functional sites in protein structures. J Mol Biol 339: 607–633.
68. Binkowski TA, Freeman P, Liang J (2004) pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. Nucleic Acids Res 32: W555–W558.
69. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM (2006) A method for localizing ligand binding pockets in protein structures. Proteins 62: 479–488.
70. Kinoshita K, Furui J, Nakamura H (2002) Identification of protein functions from a molecular surface database, eF-site. J Struct Funct Genomics 2: 9–22.
71. Schmitt S, Kuhn D, Klebe G (2002) A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 323: 387–406.
72. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. Nucleic Acids Res 32: W549–W554.
73. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA (2005) PDBSite: a database of the 3D structure of protein functional sites. Nucleic Acids Res 33: D183–D187.
74. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33: W89–W93.

Keywords:
bioinformatics, structure, sequence, protein design

Competing Interests:
The authors have declared that no competing interests exist.

75. Wallace A, Borkakoti N, Thornton J (1997) TESS: A Geometric Hashing Algorithm for Deriving 3D Coordinate Templates for Searching Structural Databases. Application to Enzyme Active Sites. Protein Sci 6: 2308–2323.
76. Kleywegt GJ (1999) Recognition of spatial motifs in protein structures. J Mol Biol 285: 1887–1897.
77. Stark A, Sunyaev S, Russell RB (2003) A model for statistical significance of local similarities in structure. J Mol Biol 326: 1307–1316.
78. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 243: 327–344.
79. Polacco BJ, Babbitt PC (2006) Automated discovery of 3D motifs for protein function annotation. Bioinformatics 22: 723–730.
80. Laskowski RA, Watson JD, Thornton JM (2005) Protein function prediction using local 3D templates. J Mol Biol 351: 614–626.
81. Redfern OC, Dessailly BH, Dallman TJ, Sillitoe I, Orengo C a (2009) FLORA: a novel method to predict protein function from structure in diverse superfamilies. PLoS Comput Biol 5: e1000485.
82. Kristensen DM, Ward RM, Lisewski AM, Erdin S, Chen BY, et al. (2008) Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. BMC Bioinformatics 9: 17.
83. Venner E, Lisewski AM, Erdin S, Ward RM, Amin SR, et al. (2010) Accurate protein structure annotation through competitive diffusion of enzymatic functions over a network of local evolutionary similarities. PLoS One 5: e14286.
84. Nemoto W, Toh H (2012) Functional region prediction with a set of appropriate homologous sequences-an index for sequence selection by integrating structure and sequence information with spatial statistics. BMC Struct Biol 12: 11.

85. Mihalek I, Res I, Lichtarge O (2006) Evolutionary and structural feedback on selection of sequences for comparative analysis of proteins. Proteins 63: 87–99.
86. Mihalek I, Res I, Lichtarge O (2006) A structure and evolution-guided Monte Carlo sequence selection strategy for multiple alignment-based analysis of proteins. Bioinformatics 22: 149–156.
87. Mihalek I, Res I, Lichtarge O (2004) A Family of Evolution-Entropy Hybrid Methods for Ranking Protein Residues by Importance. J Mol Biol 336: 1265–1282.
88. Abhiman S, Sonnhammer ELL (2005) FunShift: a database of function shift analysis on protein subfamilies. Nucleic Acids Res 33: D197–D200.
89. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Res 40: D290–301.
90. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129–2141.
91. Lee D a, Rentzsch R, Orengo C (2010) GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. Nucleic Acids Res 38: 720–737.
92. Sjölander K (1998) Phylogenetic inference in protein superfamilies: analysis of SH2 domains. Proc Int Conf Intell Syst Mol Biol ISMB Int Conf Intell Syst Mol Biol 6: 165–174.
93. Bazzoli A, Tettamanzi AGB, Zhang Y (2011) Computational protein design and large-scale assessment by I-TASSER structure assembly simulations. J Mol Biol 407: 764–776.
94. Mitra P, Shultis D, Zhang Y (2013) EvoDesign : de novo protein design based on structural and evolutionary profiles. 41: 273–280.