COEVOLVED RESIDUES AND THE FUNCTIONAL ASSOCIATION FOR INTRINSICALLY DISORDERED PROTEINS

CHAN-SEOK JEONG and DONGSUP KIM*

Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST)
335 Gwahangno, Yuseong-gu, Daejeon 305-701, Republic of Korea
**E-mail: kds@kaist.ac.kr*

The evolution of intrinsically disordered proteins has been studied primarily by focusing on evolutionary changes at an individual position such as substitution and conservation, but the evolutionary association between disordered residues has not been comprehensively investigated. Here, we analyze the distribution of residue-residue coevolution for disordered proteins. We reveal that the degree of coevolved residues significantly decreases in disordered regions regardless of the sequence propensity, and the degree distribution of coevolved and conserved residues exclusively differs in each functional category. Consequently, the coevolution information can be useful for predicting intrinsic disorder and understanding biological functions of a disordered region from the sequence.

1. Introduction

Intrinsically disordered proteins represent proteins that do not have a well-structured fold and undergo dynamic conformational changes. As disordered proteins perform important and complex biological functions such as signaling, regulation, and post-translational modification,[1,2] understanding disordered proteins becomes increasingly important. However, our knowledge on the evolution of disordered proteins and its relationship with functions have been limited.

Because of the absence of well-defined structures, sequence-based analysis plays a crucial role in predicting disordered proteins and the biological functions.[3] Initially, a low sequence complexity has been indicated as a feature of highly disordered proteins.[4] A study that analyzed sequence alignments of disordered proteins has indicated that disordered proteins have different evolutionary patterns of substitutions from ordered proteins.[5] Moreover, a model-based study has revealed that the likelihood of evolutionary mutations and the conservative substitutions for disordered proteins are different from those for ordered proteins.[6] Furthermore, an *in silico* mutagenesis study has compared the conservation of secondary structures such as helices and strands and intrinsic disorders.[7] Although the *in silico* study is based on order-biased evolutionary models,[8] it has shown that long intrinsic disorders are not easily maintained and requires a specific effort for the disorder conservation while the secondary structures are tolerant to random mutations. Therefore, the evolution of disordered proteins exhibits large differences in sequence complexity, amino acid composition, substitution, and conservation.

While the distribution of amino acids at an individual residue has been extensively studied, the evolutionary association between distinct residues, called herein coevolution, has not been explored sufficiently. Coevolution is useful information that complements conservation information, because it reveals inter-relationship between variable residues and differentiate the associated residues from other independently evolved residues. A number of studies have shown that coevolution information is related with structural and functional knowledge.[9–14]

However, the coevolution of disordered protein has not been comprehensively investigated. Although a recent study have indicated that the degree of coevolution in disordered proteins is low because of the rapid evolution rate,[14] the relationship between the coevolution signal and functionality of disordered proteins has not been investigated. The use of coevolution information for disordered proteins can be useful, because conservation information only shows a limited performance in associating the evolutionary information with the functions of disordered proteins. Particularly, according to a recent study where intrinsic disorders are classified with the evolutionary conservation of disorder and sequence, a majority of canonical intrinsic disorders belong to the "flexible disorder" category where the disorder is evolutionarily conserved but the sequences are not.[15] Consequently, the functions frequently occurred with disordered regions are not differentiated solely by the sequence conservation information, and an additional information compensating the sequence conservation is required to understand the functions.

In this study, we investigate the distribution of coevolved residues in disordered proteins experimentally determined. When comparing to non-disordered regions, disordered regions show a low degree of coevolved residues. Additionally, we investigate the relationship between coevolution and biological functions of evolutionary conserved disorders in combination with conservation information, and reveal that each functional category has a unique degree distribution of coevolved and conserved residues.

2. Materials and methods

2.1. *Data set*

Disordered proteins are collected from DisProt v5.7,[16] a manually curated database with experimentally determined structural disorders. To prevent amino acid sequence bias, the disordered proteins are clustered by running CD-HIT[17] with the maximum sequence identity threshold of <60%, and only the representative proteins are taken. Next, the proteins with less than ten homolog sequences in the multiple sequence alignments, and the residues with less than five sequences aligned at the positions are excluded for a reliable estimation of evolutionary information. Consequently, 519 disordered proteins with 57,261 disordered and 163,470 non-disordered residues are considered.

For comparison, ordered protein set consisting of 83 protein with 17,451 residues is used. The proteins have well-determined structures and exclusively represent distinct protein families. The proteins are originally organized and used for evaluating the performance of correlated mutation algorithms that estimating coevolution.[11,18,19]

2.2. *Multiple sequence alignment construction*

For each disordered or ordered protein, the multiple sequence alignment is automatically constructed with the homolog sequences. Homolog sequences are collected by running PSI-BLAST[20] with the options "-e 0.001 -h 0.001 -j 3" against the NCBI NR database clustered with <90% sequence identity, and then the hits with <50% alignment coverage are omitted by running the script program, alignhits.pl in the HHsearch package,[21] with the options "-e 0.01 -cov 50.". Subsequently, the multiple sequence alignment is constructed by running

MUSCLE[22] on the remaining hits with the fastest option "-stable -quiet -maxiters 1 -diags -sv -distance1 kbit20_3."

2.3. *Coevolution estimation*

A coevolution score between two residues is estimated with MIp measurement.[19] The MIp score is derived from the mutual information score defined as

$$\begin{aligned}\mathrm{MI}(i,j) &= \sum_x \sum_y P(x_i, y_j) \log\left(\frac{P(x_i, y_j)}{P(x_i)\, P(y_j)}\right)\\ P(x_i) &= \sum_y P(x_i, y_j)\\ P(y_j) &= \sum_x P(x_i, y_j),\end{aligned}$$

where $P(x_i, y_j)$ is the joint probability of amino acid x being at position i and amino acid y being at position j. The MIp score is calculated by removing the background noise from the original mutual information score as follows,

$$\mathrm{MIp}(i,j) = \mathrm{MI}(i,j) - \frac{\mathrm{MI}(i,\cdot)\mathrm{MI}(\cdot,j)}{\mathrm{MI}(\cdot,\cdot)},$$

where $\mathrm{MI}(i,\cdot)$ and $\mathrm{MI}(\cdot,j)$ is the mutual information score at position i and position j averaged over all other positions, respectively, and $\mathrm{MI}(\cdot,\cdot)$ is the MI value averaged over all pairs of positions. To prevent an erroneous measurement and achieve an accurate estimation of coevolution signal, we incorporate the pseudocount and marginal probability constraint based on sequence profile. To address the sequence redundancy of multiple sequence alignment, the sequence weight and the effective number of sequences are calculated using the position-based sequence weight[23] and the exponential of negative entropy averaged over all columns,[24] respectively. The above procedure is available via web [a].

The estimated MIp scores are standardized into Z-scores, and the residue pairs whose MIp Z-scores exceed 4.0 are chosen as coevolved residue pairs. Finally, coevolved residues are defined as the residues comprising coevolved residue pairs.

2.4. *Sequence conservation estimation*

Sequence conservation is estimated as Shannon's entropy with the reverse sign.[25] For example, a conservation score at position i is defined as

$$\mathrm{C}(x_i) = \sum_x P(x_i) \log\left(P(x_i)\right),$$

where $P(x_i)$ is the probability of amino acid x being at position i. Similar to the coevolution estimation, the same methods for estimating the sequence weight and the effective number of sequences are also used in the conservation score estimation.

The entropy-based conservation scores are standardized into Z-scores, and the residues whose Z-score exceed 0.8 are chosen as conserved residues.

[a] http://binfolab12.kaist.ac.kr/cmat/

2.5. *Disorder conservation estimation*

The degree of conservation of structural disorder is estimated by the ratio of predicted disordered residues among the residues aligned at the position as following the previous study.[15] For each sequence in multiple sequence alignment, the intrinsic disorders are predicted by running DisEMBL.[26] If a residue is experimentally determined as a disordered residue, and more than 40% of the aligned sequence positions are predicted as disordered residues, the residue is defined as a conserved disordered residue.

2.6. *Functional categories*

To investigate the relationship between the evolutionary information and biological functions of disordered regions, we incorporate the functional categories annotated in DisProt. There exist three kinds of categories, called structural functional type, functional class, and functional subclass, respectively.[2,27–29] The structural functional type represents a structural transition upon performing the function, and the functional class and subclass represent biological functions at different levels of details. As the disorder functions are not categorized exclusively, a disordered region can belong to multiple functional classes and subclasses, and subsequently affect their statistics simultaneously. At functional subclass level, we only consider the functional subclasses assigned at more than 1,000 distinct residues.

3. Results

3.1. *Distribution of coevolved residues for disordered proteins*

We found that the fraction of coevolved residues differs in disordered regions. The fraction of coevolved residues for disordered proteins is 36.02%, similar to that for ordered proteins, 37.12%. However, separating disorder and non-disordered regions in disordered proteins, the fraction of coevolved residues differs as 24.92% and 36.19%, respectively. That is, the degree of coevolved residues significantly drops in disordered regions. As the disordered proteins have different sequence lengths, we also examined the the degree of coevolution with respect to the degree of disorder. Figure 1 shows the frequencies of the different ratios of disordered and coevolved residues for disordered proteins. As the proteins have more disordered residues, the average fractions of coevolved residues decrease. Thus, the degree of coevolution for the disordered proteins is negatively correlated with the degree of intrinsic disorder.

Since the coevolution scores are estimated from multiple sequence alignments, the sequence diversity and the sequence complexity may influence the different degree of coevolution signals between disordered and ordered proteins, and the alignment quality also can. To check the effect of sequence diversity, we first examined the fraction of coevolved residues against disordered and non-disordered residues with the number of aligned sequences and another sequence diversity measure, the effective number of aligned sequences defined as the exponential of negative entropy averaged over all columns of aligned sequences.[24] As shown in Figure 2A, the fraction of coevolved residues for disordered residues is less than that for non-disordered residues. To take account of sequence redundancy, the degrees of coevolution of the disordered and non-disordered residues are compared after binning them according to the effective number of sequences. As shown in Figure 2B, the disordered residues show smaller fractions of coevolved residues, indicating that disordered residues have a lower degree of coevolution signal than non-disordered residues regardless of the diversity of the homolog sequences. Second, we examined the fraction of coevolved residues against disordered and non-disordered residues with various range of sequence complexities as shown in Figure 2C. It has been known that disordered regions exhibit a low sequence complexity, so the different population of coevolved residues may be a result of different sequence complexity. The sequence complexity is calculated by Shannon's entropy measure with a 45-aa window as following the previous study.[4] At every range of sequence complexity, disordered residues have a smaller fraction of coevolved residues than non-disordered residues. Third, we examined the fraction of coevolved residues against disordered and non-disordered residues with various range of sequence conservation score as shown in Figure 2D. As the alignment quality is usually estimated in comparison with structural alignment, the alignment quality for disordered region is hard to estimate. Instead, sequence conservation is used as an indirect measure of alignment quality, because aligned position with high sequence conservation are more likely accurate. The fraction of coevolved residues for disordered residues is consistently less than that for non-disordered residues at every range of sequence conservation score. Consequently, the low degree of coevolution signals in disordered regions can be though as a feature of intrinsic disorder not the bias from sequence pool and alignment error.

We also investigated the degree of coevolved residues regarding to the evolutionary classification of intrinsic disorder introduced by a recent study.[15] According to the study, intrinsic disorder can be classified into three different groups, such as constrained, flexible, and non-conserved disorders. The constrained disorder consists of residues whose disorders and amino acids are evolutionarily conserved, the flexible disorder consists of residues whose disorders are evolutionarily conserved but amino acids are not, and the non-conserved disorder consists of residues whose disorders are not evolutionarily conserved. The two conserved disorders, the constrained and flexible, show the relevant association with biological functions of intrinsic disorder, while the non-conserved disorder does not show a common functional feature. When

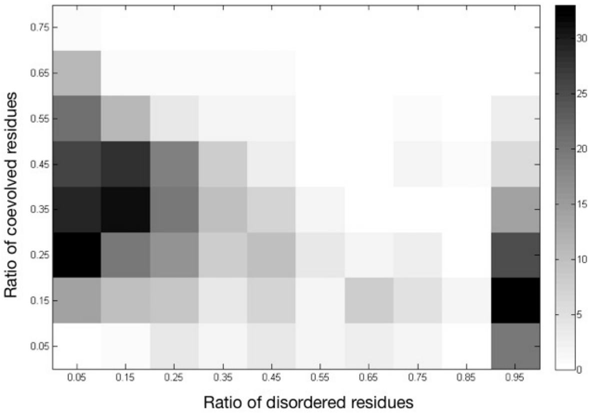

Fig. 1. Densities of disordered proteins with different ratios of disordered and coevolved residues. The intensity represents the frequency of disordered proteins with the ratios of disordered and coevolved residues.

we classified disordered residues according to the the disorder and sequence conservations, the fractions of coevolved residues for the constrained, flexible, and non-conserved disorders appear as 0.2808, 0.2536, and 0.2717, respectively. The flexible disorders show the smaller fraction of coevolved residues than the constrained and non-conserved disorders with the p-values of 2.2×10^{-5} and 6.5×10^{-3}, respectively. We repeated the evolutionary classification by replacing the sequence conservation measure with the maximum frequency of amino acids at aligned sequence positions as following the previous study, but the result does not change much. The fractions of coevolved residues for the constrained, flexible, and non-conserved disorders appear as 0.2799, 0.2457, and 0.2717, respectively, and the p-values of the flexible disorders against the constrained and non-conserved disorders are 4.4×10^{-17} and 1.8×10^{-4}, respectively. The previous study has indicated that flexible and constrained disorders are different in their functional characteristics, as flexible disorder usually presents short linear motifs and performs in signaling,[30] while constrained disorder is associated with the functions requiring tight sequence constraint.[15] Moreover, it has been suggested that contained disorder can often adopt more strict conformations than flexible disorder, so its higher degree of coevolution may come from the same constraint. Another high degree of coevolution observed in non-conserved disorder can be explained with the high rate of ordered protein residue. The aligned protein residues in the non-conserved disorder possibly have a high ratio of structured residues, and thus the background coevolution signal may influence the high degree of coevolved residue in the regions. It is important to note the evolutionarily different disorder groups have distinct functionalities, and in our results, the degrees of coevolved residues differ in those groups.

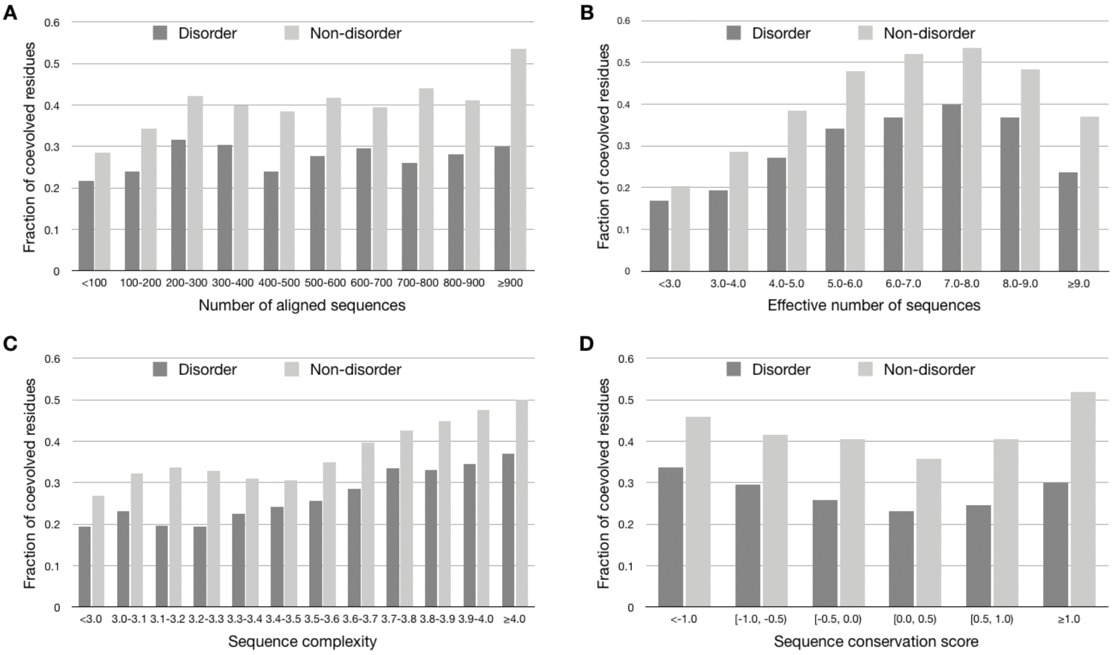

Fig. 2. Degree distribution of coevolved residues in disordered and non-disordered regions. (A) Fraction of coevolved residues to all residues with different ranges of number of aligned sequences. (B) Fraction of coevolved residues to all residues with different ranges of sequence complexity. (C) Fraction of coevolved residues to all residues with different ranges of sequence conservation score.

That is, the coevolution information can be a useful for understanding biological functions of intrinsic disorder.

3.2. *Relationship between coevolution and functions*

Intrinsic disorders have been classified with the evolutionary conservation in the predicted disorder and the sequence, and evolutionary conserved disordered regions have the frequently occurred functions of intrinsic disorders.[15] However, major functions of intrinsic disorder such as signaling and regulation are not differentiated only by the sequence conservation information. Here, we further investigate the relationship between coevolution information and common functions of evolutionary conserved disorders in combination with sequence conservation information.

The content of conserved and coevolved residues significantly differs according to the structural functional type. As shown in Figure 3, two main structural functional types, the disorder that performs the function in disordered state and the disorder-to-order that undergoes structural transition from disordered to ordered state, have distinct frequencies of conserved and coevolved residues. The disorder-to-order type has more conserved and coevolved residues than disorder type, which implies the higher structural constraints upon the structural transition. Additionally, this can support the association between constrained disorder and the structural transition that the previous study has speculated.[15] On the other hand, the disorders whose functions arise within the disordered state are less evolutionarily restrained. As consequence, the structural functional types have distinct proportion of evolutionary conserved and coevolved residues.

Each functional class differs in the ratio of conserved and coevolved residues as shown in Figure 4. The effector class with permanent binding, that recognizes a partner molecule and modifies the activity, shows the highest degree of coevolved residues among the six functional classes. Another class with permanent binding, the assembly class, have relatively more conserved residues than the effector class. The other class with permanent binding, the scavenger class, has the highest ratio of conserved residues. Since the effector and assembly classes have known as related to disorder-to-order transitions,[31–33] their high degrees of conservation and coevolution reflect the functionality. However, the preferred contents are different in combination of both evolutionary information. Specifically, the effector class prefers the coevolved variable residues most. In the similar aspect, the modification site and the chaperone classes share transient binding property,[2,27] but are distinct from each other with the degree of conser-

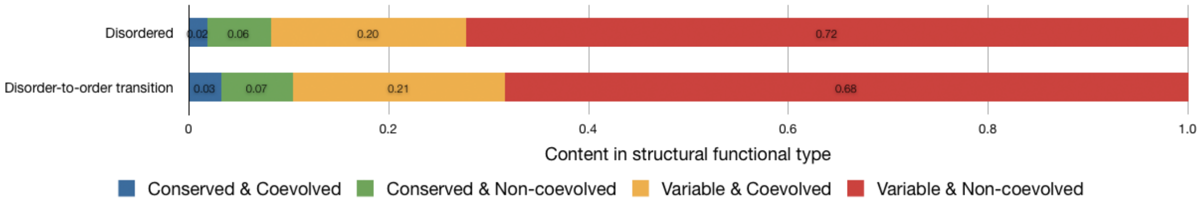

Fig. 3. Relative frequency of conserved and coevolved residues in various structural functional types of disordered residues.

vation and coevolution. The modification site class prefers conservation, while the chaperone class does coevolution of variable residues. The entropic chain class has a large fraction of variable and non-coevolved residues, which reflects the functionality arising from the disordered state.[2] The modification site and the entropic chain classes share a similar content of conservation and coevolution information. Nevertheless, the use of coevolution information reasonably differentiates the effector, the assembly, and the modification site classes, which have almost similar proportions of conserved residues. That is, the degree of coevolution in each functional class can reflect the function and uniquely discriminate the relevant functional class in combination with the degree of conservation.

At the functional subclass level, each functional subclass has a distinct ratio of conserved and coevolved residues as shown in Figure 5. The protein-DNA binding subclass has the highest ratio of conserved residues. The substrate/ligand binding, intra protein interaction, and transactivation subclasses commonly have a high degrees of conserved residues, but significantly differ from each other with the fractions of coevolved residues. In particular, the substrate/ligand binding subclass has a high fraction of coevolved residues, and the other two subclasses have low fractions of coevolved residues. Importantly, various kinds of binding subclasses, such as the protein-protein, protein-DNA, protein-rRNA, substrate/ligand, and metal binding subclasses, are meaningfully separated with the combination of the ratios of coevolved and conserved residues, despite their similarity with permanent binding.[2,27,31–33] The phosphorylation subclass, that comprises a majority of the modification site class, and the autoregulatory subclass, that frequently undergoes some kinds of modifications during playing the role,[2] similarly have a large fraction of variable and non-coevolved residues, but their frequencies conserved and coevolved residues separate the functional subclasses. As at the functional class level, some functional subclasses such as the metal binding, the protein-protein binding, and the phosphorylation subclasses are differentiated by combining coevolution information, even though they have a common enrichment of conserved residues. Therefore, the functional subclasses can be uniquely separated by the degree of coevolution and conservation, likewise the functional classes.

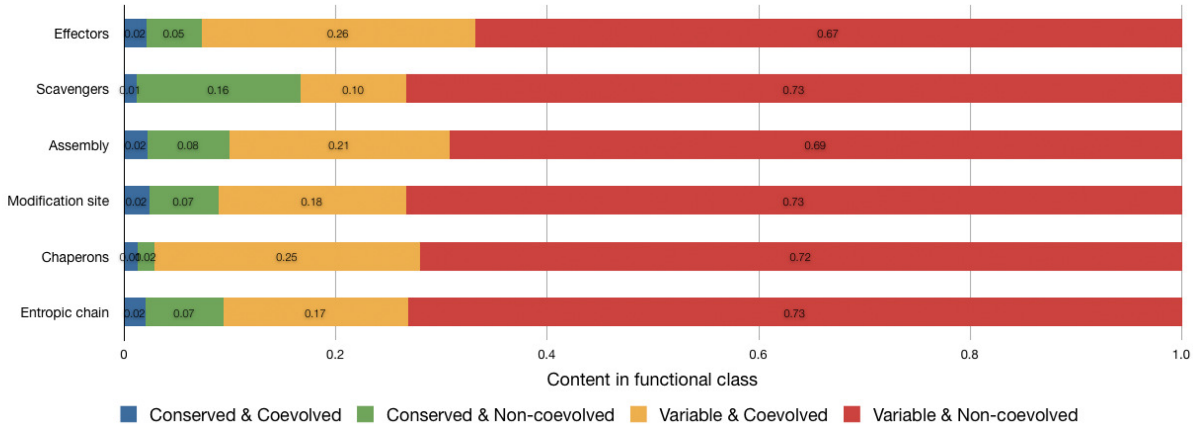

Fig. 4. Relative frequency of conserved and coevolved residues in various functional classes of disordered residues.

4. Discussion

We have demonstrated that coevolved residues are less populated in disordered region. The lack of structural constraints for disordered regions would be the main reason for the degree distribution, because of two reasons. The first reason is that the structural constraints such as intra-molecular contacts are known as one of main factors that develop coevolution signals,[34,35] and the second reason is that the degree of coevolution in disordered regions shows a negative correlation with the degree of disordered residues. That is, unlike ordered residues, disordered residues do not lie in strong structural constraints, lacking a well-structured fold and dynamically changing the conformations, which consequently reduces the degree of evolutionarily coupled residues.

The low degree of coevolved residues in disordered regions can be an effective indicator for disorder prediction. To explore the usefulness of coevolution information for disorder prediction, the enrichment of coevolved residues in neighboring residues within ±7 aa long has been estimated as shown in Figure 6. For disordered residues, 44.4% of disordered residues have zero or one coevolved residues among the neighboring residues. On the other hand, for ordered residues, the number of coevolved residues are evenly distributed regardless of the secondary structure elements, as 5.6–18.0%. Hence, the coevolution information can be useful for disorder prediction. Although some of currently available disorder prediction methods, categorized to as contact-based method, implicitly incorporate coevolution information, those methods do not fully exploit the effectiveness.[36] Some of them just use contact potential without evolutionary information of a target protein.[37,38] Even though Ucons,[39] that directly incorporates contact predictions using evolutionary information,[40] coevolutionary information are not included in the contact prediction procedure. Therefore, the use of coevolutionary information

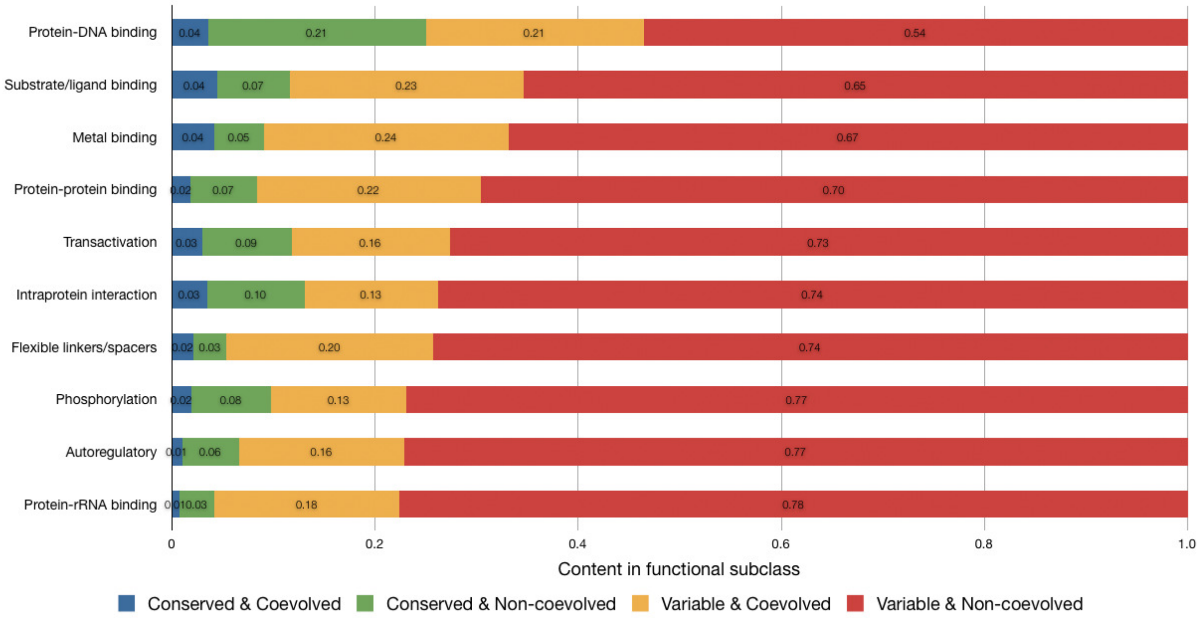

Fig. 5. Relative frequency of conserved and coevolved residues in various functional subclasses of disordered residues.

can effectively improve the current methods for disorder prediction.

Even though disordered residues tend to be less coevolved than ordered residues, a relative ratio of coevolved residues in a disordered region can be an effective indicator to the biological function. When we have examined the proportions of conserved and coevolved residues for various functional annotations at different levels and criteria, the functional annotations have exclusively distinct proportional features. In particularly, many of those functional annotations that share a similar sequence conservation degree can be differently classified in combination with coevolution information. Therefore, the coevolution information can be useful for specifying biological functions of disordered proteins as well as the disorder prediction.

According to Jeon et. al.'s study[14] that explores the relationship between coevolution and conformation changes, the degree of coevolution in flexible regions of ordered proteins encodes the structural transitions. In addition, they have concluded that disordered proteins have a low degree of coevolution, not reflecting the high flexibility. However, in the present study, we have compared the degree distributions of coevolved residues to the overall distribution as background statistics, and revealed the association between coevolution and dynamic behavior such as structural functional types. Accordingly, structural transitions of disordered proteins seem to be encoded as the coevolution signals like those of ordered proteins, but additional studies are needed to clarify the functional association. Because disordered and ordered proteins have different background degrees of coevolution, prior knowledge for the background coevolution upon the structural disorder need to be investigated and incorporated to normalize the coevolution signal, which can be critical to revealing the functional roles of coevolution information.[41]

Coevolution information can be useful for understanding disordered proteins due to three reasons. First, as we have demonstrated, the degree of coevolved residues is related with functional classifications. For example, disordered residues whose function arises from the disordered state show a low degree of coevolution signal, and, on the other hand, disordered residues whose function requires specific recognitions as in effector class show a high degree

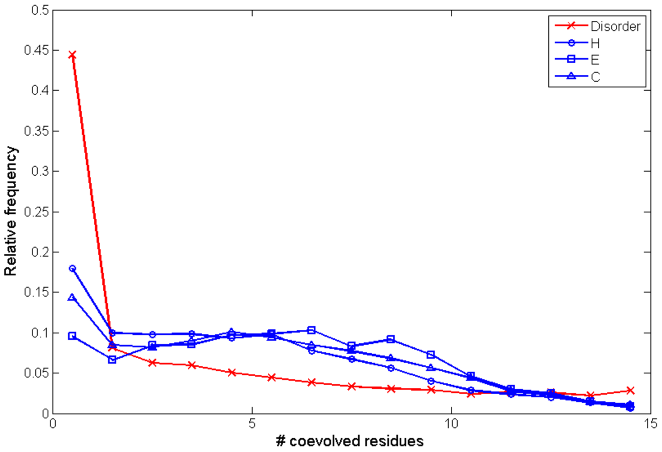

Fig. 6. Distribution of the number of coevolved residues within neighboring positions for disordered and ordered residues. Disorder indicates disordered residues. H, E, and C indicate ordered residues belonging to helix, sheet, and coil by secondary structure, respectively.

of coevolution signal. Second, the coevolution information successfully compensates the conservation information. We have demonstrated a group of functional categories, similar to each other with the degree of conservation but different with the degrees of coevolution signals. Third, coevolution information describes evolutionary association between residues which cannot be represented by conventional sequence analyses. Thus, coevolution analysis would be beneficial to understanding the evolution of disordered proteins.

Although we have incorporated an advanced coevolution estimate with high reliability and robustness, the analysis results are still dependent on multiple sequence alignments. Because the alignment quality is usually estimated in comparison with structural alignment and most of alignment methods are designed for ordered proteins,[42] alignment quality of disordered proteins are difficult to be estimated. Even though, in this study, we have used sequence conservation score as an indirect measure of alignment quality, the effect of alignment quality for disordered proteins needs to be examined more apparently. As a separately calculated substitution matrix for disorder proteins has improved the alignments of them, similarly, a development of multiple sequence alignment methods for disordered proteins would be useful for an accurate coevolution analysis of them.

This study can be extended to a genome-scale analysis by combining disorder prediction methodologies.[43] In the present work, we have chosen only the experimentally determined disordered proteins from DisProt database. Although the sequence redundancy has been removed with sequence identity, the result statistics still depends on the content of the original database. Thus, by incorporating disorder prediction methodologies and then applying the present analysis procedures, more realistic knowledge about coevolution in disordered proteins can be obtained as a lot of prediction-based studies.[15,44–46] Moreover, various kinds of biological information associated with intrinsic disorder can be easily combined and considered in the analysis procedure.

Acknowledgments

We thank all Bioinformatics and Computational Biology Laboratory (BCBL) members for helpful discussion. This work was supported by a grant of the Korea Healthcare technology R&D Project, Ministry for Health, Welfare & Family Affairs, Republic of Korea [A092006]; and the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MEST) [2009-0086964].

References

1. P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky and A. K. Dunker, *Biophys. J.* **92**, 1439 (March 2007).
2. A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradovic, *Biochemistry* **41**, 6573 (May 2002).
3. Z. Dosztanyi, B. Meszaros and I. Simon, *Brief Bioinformatics* **11**, 225 (March 2010).
4. P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown and A. K. Dunker, *Proteins* **42**, 38 (January 2001).
5. P. Radivojac, Z. Obradovic, C. J. Brown and A. K. Dunker, *Pac Symp Biocomput* , 589 (2002).
6. C. J. Brown, A. K. Johnson and G. W. Daughdrill, *Mol Biol Evol* **27**, 609 (March 2010).

7. C. Schaefer, A. Schlessinger and B. Rost, *Bioinformatics* **26**, 625 (March 2010).
8. C. J. Brown, A. K. Johnson, A. K. Dunker and G. W. Daughdrill, *Curr Opin Struct Biol* **21**, 441 (June 2011).
9. S. Chakrabarti and A. R. Panchenko, *PLoS ONE* **5**, p. e8591 (January 2010).
10. A. Kowarsch, A. Fuchs, D. Frishman and P. Pagel, *PLoS Comput Biol* **6**, p. e1000923 (2010).
11. B.-C. Lee and D. Kim, *Bioinformatics* **25**, 2506 (October 2009).
12. N. Halabi, O. Rivoire, S. Leibler and R. Ranganathan, *Cell* **138**, 774 (August 2009).
13. S. Chakrabarti and A. R. Panchenko, *Proteins* **75**, 231 (April 2009).
14. J. Jeon, H.-J. Nam, Y. Choi, J.-S. Yang, J. Hwang and S. Kim, *Mol Biol Evol* (April 2011).
15. J. Bellay, S. Han, M. Michaut, T. Kim, M. Costanzo, B. J. Andrews, C. Boone, G. D. Bader, C. L. Myers and P. M. Kim, *Genome biology* **12**, p. R14 (February 2011).
16. M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic and A. K. Dunker, *Nucleic Acids Res* **35**, D786 (January 2007).
17. W. Li and A. Godzik, *Bioinformatics* **22**, 1658 (July 2006).
18. C. M. Buslje, J. Santos, J. M. Delfino and M. Nielsen, *Bioinformatics* **25**, 1125 (May 2009).
19. S. D. Dunn, L. M. Wahl and G. B. Gloor, *Bioinformatics* **24**, 333 (February 2008).
20. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res* **25**, 3389 (September 1997).
21. J. Söding, *Bioinformatics* **21**, 951 (April 2005).
22. R. C. Edgar, *Nucleic Acids Res* **32**, 1792 (January 2004).
23. S. Henikoff and J. G. Henikoff, *J Mol Biol* **243**, 574 (November 1994).
24. A. Biegert and J. Söding, *Proc Natl Acad Sci USA* **106**, 3770 (March 2009).
25. J. Pei and N. V. Grishin, *Bioinformatics* **17**, 700 (August 2001).
26. R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson and R. B. Russell, *Structure/Folding and Design* **11**, 1453 (November 2003).
27. P. Tompa, *Trends Biochem. Sci.* **27**, 527 (October 2002).
28. V. N. Uversky, *Protein Sci* **11**, 739 (April 2002).
29. H. J. Dyson and P. E. Wright, *Nat. Rev. Mol. Cell Biol.* **6**, 197 (March 2005).
30. F. Diella, N. Haslam, C. Chica, A. Budd, S. Michael, N. P. Brown, G. Trave and T. J. Gibson, *Front. Biosci.* **13**, 6580 (2008).
31. A. P. Demchenko, *J. Mol. Recognit.* **14**, 42 (2001).
32. H. J. Dyson and P. E. Wright, *Curr Opin Struct Biol* **12**, 54 (February 2002).
33. R. S. Spolar and M. T. Record, *Science* **263**, 777 (February 1994).
34. G. Shackelford and K. Karplus, *Proteins* **69 Suppl 8**, 159 (January 2007).
35. U. Göbel, C. Sander, R. Schneider and A. Valencia, *Proteins* **18**, 309 (April 1994).
36. F. Orosz and J. Ovádi, *Bioinformatics* **27**, 1449 (June 2011).
37. Z. Dosztanyi, V. Csizmok, P. Tompa and I. Simon, *Bioinformatics* **21**, 3433 (August 2005).
38. O. V. Galzitskaya, S. O. Garbuzynskiy and M. Y. Lobanov, *Bioinformatics* **22**, 2948 (December 2006).
39. A. Schlessinger, M. Punta and B. Rost, *Bioinformatics* **23**, 2376 (September 2007).
40. M. Punta and B. Rost, *Bioinformatics* **21**, 2960 (July 2005).
41. A. D. Fernandes and G. B. Gloor, *Bioinformatics* **26**, 1135 (May 2010).
42. J. D. Thompson, B. Linard, O. Lecompte and O. Poch, *PLoS ONE* **6**, p. e18093 (2011).
43. F. Ferron, S. Longhi, B. Canard and D. Karlin, *Proteins* **65**, 1 (October 2006).
44. J. W. Chen, P. Romero, V. N. Uversky and A. K. Dunker, *J. Proteome Res.* **5**, 888 (April 2006).
45. J. W. Chen, P. Romero, V. N. Uversky and A. K. Dunker, *J. Proteome Res.* **5**, 879 (April 2006).
46. J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones, *J Mol Biol* **337**, 635 (March 2004).