A NOVEL PHYLOGENY-DEPENDENT COEVOLUTION ALGORITHM

Nurdan KURU[1] and Ogün ADEBALİ[1,2]

[1] Faculty of Engineering and Natural Sciences, Sabancı University
Orta Mahalle, 34956, Tuzla, İstanbul, Turkey
[2] TÜBİTAK Research Institute for Fundamental Sciences, 41470 Gebze, Turkey
phone: + (216)5687043, email: nurdan.kuru@sabanciuniv.edu, ogun.adebali@sabanciuniv.edu

ABSTRACT

Genetic elements that work together evolve together. Co-evolution trends of amino acids observed within or between genes provide important information about many features of proteins, from structure to function. These trends are of interest to the diagnosis of genetic diseases in humans. Most coevolution methods are based solely on multiple sequence alignments and ignore phylogenetic relatedness, thus a shared evolutionary history. This incomplete information hinders the accurate identification of co-evolving protein regions. There is a need for a high-accuracy algorithm that yields better sensitivity with fewer false positives. Such a tool would help us better understand not only coevolution but also many related biological aspects such as disease diagnosis, molecular functions of different protein families, and mutations that are effective in the development of drug resistance.

In this study, we hypothesize that a phylogeny-derived coevolution algorithm supported with ancestral reconstruction yields more accurate coevolution signals compared to the methods based on multiple sequence alignments. We present a novel approach based on substitution mapping of amino acid changes onto the phylogenetic tree. Our method evaluates the coevolving positions not only by considering the total change but also whether the changes are repeated or not, the effect of the amino acids in the position in terms of physicochemical properties, and the conservation of the positions. For each leaf of the phylogenetic tree, we travel through the nodes and compute the total amount of substitution per branch based on the probability differences of ancestral amino acids between neighboring nodes in the tree. The sequences with gaps or showing inconsistency with the remaining sequences around the position in question are eliminated in score computation. In other words, we remove the false coevolution signals that are resulted from unaligned regions by trimming corresponding sequences from the phylogenetic tree. Each position pair is scored between 0 and 1 per leaf by accounting for parallel changes observed on similar branches. If a coevolution signal exists for at least 20% of the leaves, we assign a score to the corresponding position pairs. With the methods we develop, we highlight independent and repetitive mutations and correctly interpret the coevolution between positions with a slow evolutionary pace and conserved positions.

We compare the performance of our algorithm against CAPS[1], CoMap[2] and iBIS2Analyzer[3]. Figure 1 shows some position pairs analyzed in terms of coevolution. In figure 1a and b, we present two position pairs that were assigned a high score by our approach. The pair in figure 1b was not detected by CAPS, CoMap and iBIS2Analyzer although the coevolution signal exists. Figure 1c shows an example position pair that was mislabelled by CAPS because of unaligned regions (shown in black rectangles). These examples illustrate that our approach better identifies the coevolving position pairs by eliminating the MSA-based problems and phylogenetic dependency that are ignored by existing tools.

REFERENCES

1. Fares, M. A., & McNally, D. (2006). CAPS: coevolution analysis using protein sequences. Bioinformatics, 22(22), 2821-2822.
2. Dutheil, J., & Galtier, N. (2007). Detecting groups of coevolving positions in a molecule: a clustering approach. BMC evolutionary biology, 7(1), 1-18.
3. Oteri, F., Sarti, E., Nadalin, F., & Carbone, A. (2022). iBIS2Analyzer: a web server for a phylogeny-driven coevolution analysis of protein families. Nucleic Acids Research.

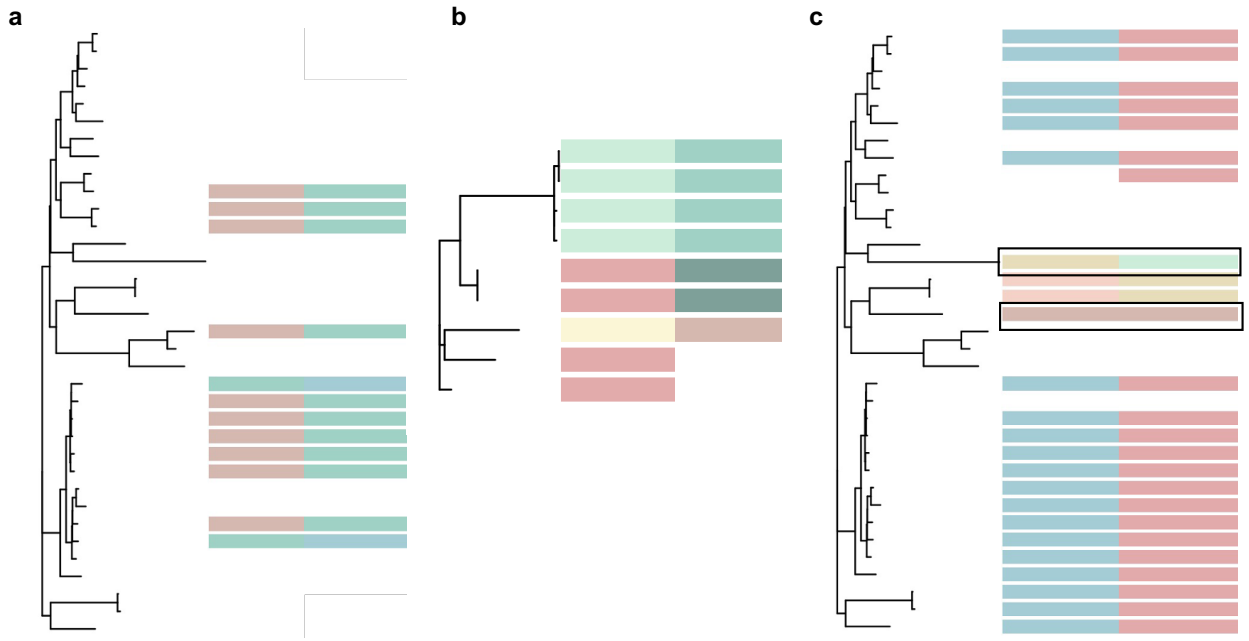

Figure 1. a, b Two position pairs were assigned a high coevolution score by our approach. c The given position pair was mislabelled as coevolving residues by other tools. At each part, different colors correspond to different amino acids on the position.