Complementary Grouping of Amino Acids based on Base-Pairing

Minseo Park and B.G. Kim

Abstract— **DNA sequence analysis is at the core of bioinformatics approaches in the study of genomes, genes and proteins. Recent studies revealed that codon (triplet nucleotides) is a useful means to detect patterns of genome sequences and is the basis of amino acids, which in turn translate to proteins. Sixty four codons are commonly grouped into an alphabet of twenty amino acids. Grouping twenty amino acids into reduced alphabets may aid in detecting patterns, modeling and designing proteins, and finding consensus sequences. It can make patterns simply. In this paper, a novel approach is used to generate a reduced alphabet of twenty amino acids. It is based on base-pairing and incorporates physical features of amino acids. The base-pairing is formulated by a complementary tree algorithm. The tree preserves the relationship among four groups of amino acids (nonpolar, polar, basic, and acidic), and yet generates fifteen amino acid groups. The resulting reduced alphabets are tested on alternative splicing patterns, which is the process related to protein functions and an important process for increasing the diversity arising from a single gene, in a genome sequence. The evaluation shows that this grouping is efficient in detecting patterns of genome sequences.**

I. INTRODUCTION

Most biological functions are mediated from proteins, which are translated from twenty amino acids. Primary features of proteins are determined by the presence and arrangements of specific amino acids. Therefore, analyzing properties of amino acids can assist in predicting functions and structures of proteins [1][2].

Triplet nucleotides form the basis of sixty four codons, which are converted into twenty amino acids. Grouping twenty amino acids into a smaller alphabet is shown to be useful [3]. For example, the number of protein blocks generated from twenty amino acids is as many as 20^{10} for a polypeptide of length 10. A reduced alphabet has been shown to be useful in designing foldable sequences of a large number of protein families [4]. Significant savings in computation times can be achieved in the analysis of genome sequences with reduced alphabets [5]. In another application, a reduced alphabet size has been used for finding consensus sequences in multiple alignments [6][7]. Reduced alphabets are used in aiding protein modeling [8]. A simple grouping of amino acids according to hydrophobic features is shown to provide positive results for protein-protein interaction [9]. Although many studies focus on polar/nonpolar features, several studies suggest that more than two groups may be more appropriate [8][10][11][12].

In this paper, we propose a new grouping of amino acids, which allows for simpler genome sequence patterns than using twenty amino acids groups. The new grouping is constructed from both complementary nature of base-pairing in codons and physical properties of side chain of amino acids. The grouping is formulated into a complementary tree algorithm. Each node in the tree has two child nodes which have complementary relationship with respect to base properties. The reduced alphabets are known to be difficult to test directly on proteins or biological events [4], and the complementary grouping is tested on alternative splicing patterns which is closely related to detecting protein functions. Test results demonstrate that the complementary grouping can be useful in analyzing alternative splicing sequence patterns.

The remainder of the paper is organized as follows. In section II, we review previous amino acid classifications and propose a new classification method which takes into account both the bases within codons and physical properties of amino acids. Construction of the complementary tree is described in section III. The tree is created by base-pairing in each base within codons. In section IV, grouping of amino acids into 15 groups is proposed. Application of the new grouping to the detection of alternative splicing of *Arabidopsis thaliana* is presented. Concluding remarks and suggested future directions are included in section V.

II. CLASSIFICATION OF AMINO ACIDS

This section introduces several classifications among physical and chemical properties of amino acids side chains. A novel classification of codons is proposed according to both physical properties for side chain of amino acids and arrangement of bases.

Amino acids have diverse features according to physical and chemical properties, functional group charge, and surface tension [13]. The twenty amino acids can be commonly catalogued into four groups including basic, acidic, hydrophilic (polar) and hydrophobic (nonpolar), according to physical features of side chains in amino acids. According to a hydrophobic feature, amino acids can further be categorized into two groups, polar or non-polar as shown in Fig. 1.

Amino acids with polar side chains can be classified into three groups, basic, acidic, and neutral (Fig. 2). Relationship among four features (polar, non-polar, acidic, and basic) are depicted in Fig. 3

The grouping of amino acids by equivalence classes can be an efficient means of expressing amino acid features, and can be used to predict the structure of proteins [8][10][11][12]. In studies examining protein-protein interactions, results using hydrophobic features are superior to those using several features simultaneously without the hydrophobic feature [9].

Minseo Park is with the Department of Computer Science, University of Massachusetts, Lowell, MA, USA mpark@cs.uml.edu

B.G. Kim is with the Department of Computer Science, University of Massachusetts, Lowell, MA, USA kim@cs.uml.edu

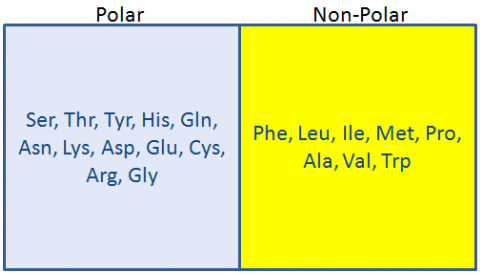

Fig. 1. Amino acid classification by polarity.

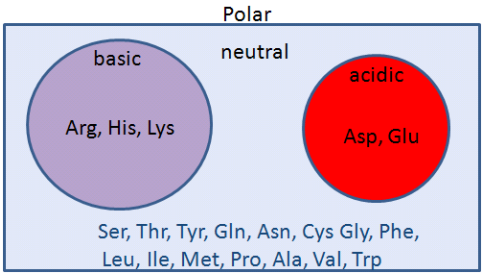

Fig. 2. Amino acid classification by physical features inside a polar group : Basic, Neutral, and Acidic.

Although grouping by polar/nonpolar features shows positive results for protein-protein interaction, several studies suggest that more than two groups might be more useful [8][10][11][12].

The new classification method proposed in this paper takes into account both arrangement of bases in codons and physical features of amino acids. We first consider the second base as the attribute for classifying codons. We then consider the first base as the next attribute for classification, followed by the third base. Such a grouping by the second codon base is shown in Fig. 4. Conventionally, codon classification is done starting with the first base as shown in Fig. 5(a). For reference, classification starting with the third base is included in Fig. 5(b). A quick inspection of the three figures reveals that groupings by the first and the third bases show more variations than the one by the second base. In fact, codons grouped according to the second base are shown to have a clear pattern, which is further elaborated into a complementary relationship in the next section.

III. COMPLEMENTARY TREE

The pattern in Fig. 4 is captured in a complementary tree. In the tree, each node has a different physical and chemical property. Each node has two child nodes, which possess a complementary relationship (base-pairing) with respect to bases.

Grouping by base-pairing is repeated until each group has one physical property for amino acids side chain. The procedure is as follows:

1) The first application of base-paring is grouping sixty four codons into two groups according to the second base such that (*U* and *A*) and (*C* and *G*), where * denotes any nucleotide among A, U, G, and C.
2) The second application of base-paring identifies fifteen codons from the (*U* and *A*) group that is classified in step 1. All of the fifteen codons have one pattern, *U*, and they are translated into four amino acids. However, they share the physical feature of being non-polar.
3) *C* within (*C* and *G*) in step 1 is partitioned into two groups, (UC* and AC*) and (CC* and GC*) according to the first base. (UC* and AC*) group has a polar feature, and (CC* and GC*) group has a non-polar property.
4) The *A* in step 2 is partitioned into (UA* and AA*) and (CA* and GA*) according to the first base. UA* and GA* are not classified any longer because each of them has the common physical property that UA* has a polar and that GA* has an acidic property.
5) The *G* is partitioned into two groups, (UG* and AG*) and (CG* and GG*) according to the first base.
6) The group (CG* and GG*) is again partitioned into two groups, CG* and GG*, by the first base. The CG* group has a basic property, and the GG* has a

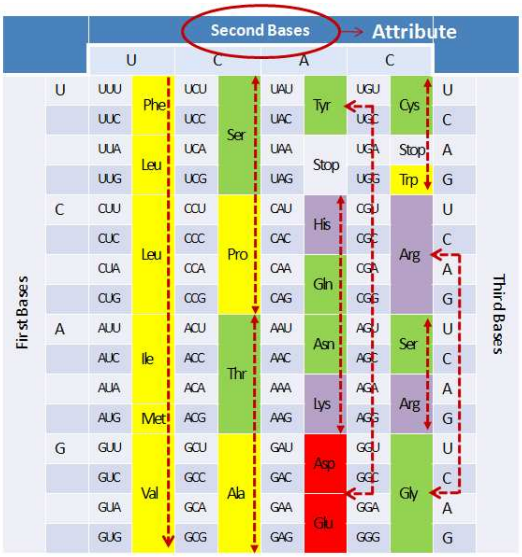

Fig. 4. Classification by the second base. The yellow indicates non-polar, green indicates polar, purple indicates basic and red indicates acidic.

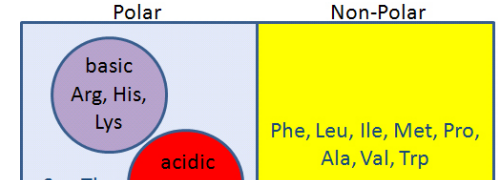

Fig. 3. The relationship among polar, non-polar, basic and acidic amino acids groups.

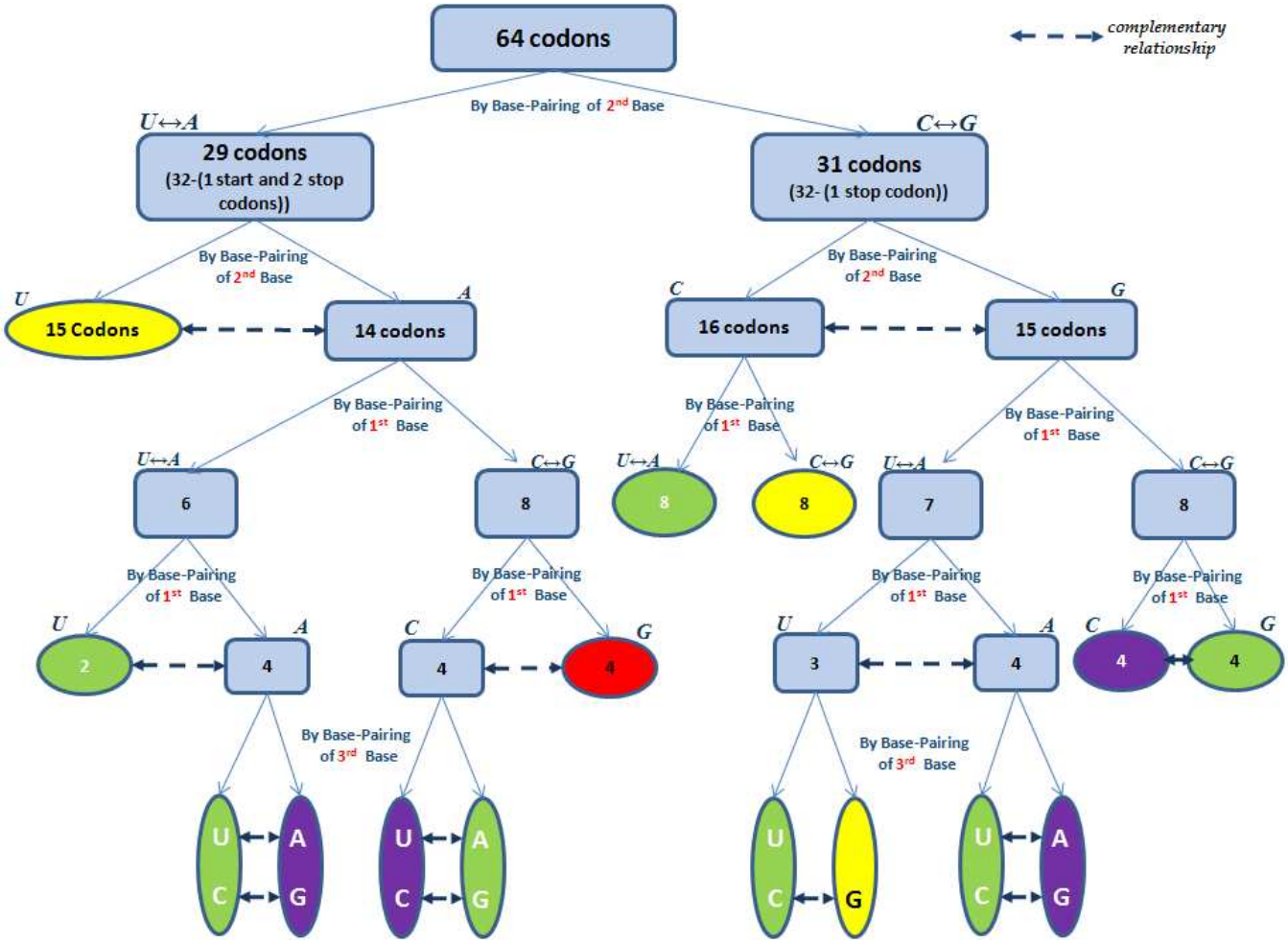

Fig. 6. The complementary tree which is created by base-pairing. The yellow indicates non-polar, green indicates polar, purple indicates basic and red indicates acidic.

Fig. 5. Classification by the first base (5(a)) and the third base (5(b)). Follow red arrows. Third-based classification of codons has more variations than first-based one.

non-polar property.

7) Remaining groups are partitioned according to the third base. Each of remaining groups has common physical features.

In Fig. 6, it can be noted that not only an opposing relationship between non-polar and polar features but also a complementary relationship between polar and basic features. Even though Fig. 2 shows that all basic features occur in the polar group, the analysis by the complementary tree shows that the two groups can have a complementary relationship.

The complementary tree also reflects physical properties of proteins. For example, fifteen codons are first filtered into one group with a non-polar feature from classification with the second bases (Fig. 6). They can be an important factor in protein-protein interaction [9]. With the complementary tree that reflects physical features and computational analysis of bases, one may be able to predict the structure and nature of proteins and patterns of genome sequences. The prediction capability allows molecular biologists to save time and conduct more precise experiments for biological events [14].

IV. Results of Various Groupings

In order to evaluate grouping methods, a series of tests is conducted on a biological event of alternative splicing, which is related to mutation of genome sequences and its translation from amino acids to proteins. The *Arabidopsis thaliana*, a flowering plant, is used as a model organism. Its

genome has been completely sequenced, is heavily annotated and is available in TAIR database [15][16]. The accuracy of alternative splicing is computed with 'Alternative Acceptor Site' section of the TIGR database [17], which uses a combination of ESTs and experimental validation for improved data quality [15][16][18]. It is already established that each chromosome of Arabidopsis has a different pattern [16].

By using fewer groups for amino acids than the traditional twenty amino acids, it is intuitively clear that detection or prediction of sequence patterns and protein functions may be less accurate. The degree of accuracy is expected to be dependent upon the alphabet size and how groupings are created. In the series of experiments, we attempt to show the results in progressively bigger alphabet sizes.

We first tested the grouping based on two physical features, polar and non-polar, and determine how well the simple grouping can detect a alternative splicing. The analysis produced poor results. Alternatively and normally spliced sites could not be distinguished with the simple grouping with the two physical features. They did not produce different patterns.

We then used the traditional four property groups, polar, non-polar, basic and acidic. Results from the grouping did not provide good results, either.

Another was grouped only by both the second bases of codons and physical features of amino acids. This method has grouped sixty four codons into nine groups. Fig. 4 shows different colors for physical features by the second bases of codons. The nine groups are as follows:

1) *The Second Base (**U**) & Non-Polar*: Phe, Leu, Ile, Met, Val

2) *The Second Base (**C**) & Polar*: Thr
3) *The Second Base (**C**) & Non-Polar*: Pro, Ala

4) *The Second Base (**A**) & Polar*: Tyr, Gln, Asn
5) *The Second Base (**A**) & Basic*: His, Lys
6) *The Second Base (**A**) & Acidic*: Asp, Gln

7) *The Second Base (**G**) & Polar*: Cys, Ser, Gly
8) *The Second Base (**G**) & Non-polar*: Trp
9) *The Second Base (**G**) & Basic*: Arg

The 9-group alphabet was used to determine spliced sites into normally and alternatively spliced sites in chromosome At1g acceptor sites of *Arabidopsis thaliana*. The result showed that only 25.9% of spliced sites were correctly classified.

Finally, the complementary grouping of 15-group alphabet was applied the same data. Mapping of sixty four codons into fifteen groups in section III is summarized below:

1) ***U***: 15 codons (Non-Polar)

2) ***UC*** *or* ***AC***: 8 codons (Polar)
3) ***CC*** *or* ***GC***: 8 codons (Non-Polar)
4) ***CG***: 4 codons (Basic)
5) ***GG***: 4 codons (Polar)

6) ***GA***: 4 codons (Acidic)

7) ***UAU*** *or* ***UAC***: 2 codons (Polar)

8) ***AAU*** *or* ***AAC***: 2 codons (Basic)
9) ***AAA*** *or* ***AAG***: 2 codons (Polar)

10) ***CAU*** *or* ***CAC***: 2 codons (Basic)
11) ***CAA*** *or* ***CAG***: 2 codons (Polar)

12) ***UGU*** *or* ***UGC***: 2 codons (Non-Polar)
13) ***UGG***: 1 codons (Polar)

14) ***AGU*** *or* ***AGC***: 2 codons (Polar)
15) ***AGA*** *or* ***AGG***: 2 codons (Basic)

The complementary grouping method provides accurate classification of 62.4% in chromosome At1g, 99.8% in chromosome At2g, 99.8% in chromosome At3g, 57.4% in chromosome At4g, and 64.4% in chromosome At5g for 'Alternative Acceptor Sites' section of TIGR, respectively. Classification result of alternatively and normally spliced sites for chromosome At3g is depicted in Fig. 7, with actual number of acceptor sites classified at each iteration of classification.

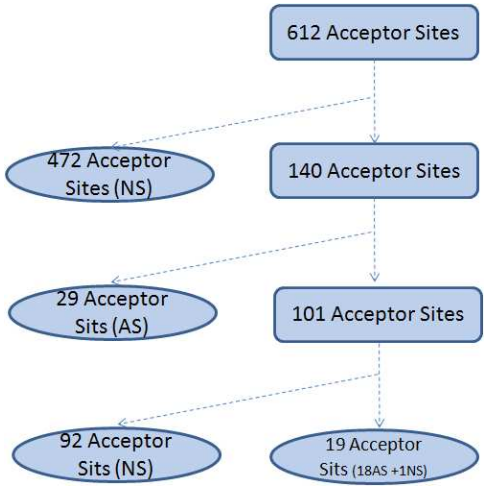

Fig. 7. Classification by fifteen groups of codons we proposed in chromosome At3g acceptor sites. NS indicates normally spliced acceptor sites, and AS indicates alternatively spliced acceptor sites.

V. CONCLUSION

In this paper, we identify the relationship between the four amino acid groups and their physical features in terms of nucleotide bases. As the relationship is established, the complementary nature of bases is seen as the primary cause of different amino acid groups. It is demonstrated that non-polar and polar features as well as basic and polar features have complementary relationships with base-pairing within

codons. The complementary relationship is captured in a tree structure, where each node represents a physical feature of amino acids. By making use of the relationship among amino acids, it is expected that detection of patterns in sequence alignments may be achieved in a simpler manner [3]. It is also expected that the proposed grouping may be useful in modeling and in protein design, and in finding consensus sequences [3][4][8] in bioinformatics approaches.

REFERENCES

[1] C. Leslie, E. Eskin, and W. Noble, "The spectrum kernel: a string kernel for SVM protein classification," in *Proceedings of the Pacific Biocomputing Symposium*, 2002, pp. 564–75.

[2] S. Martin, D. Roe, and J. Faulon, "Predicting protein-protein interactions using signature products," *Bioinformatics*, vol. 21, pp. 218–226, 2005.

[3] N. Cannata, S. Toppo, C. Romualdi, and G. Valle, "Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices," *Bioinformatics*, vol. 18, no. 8, pp. 1102–1108, 2002.

[4] L. Murphy, A. Wallqvist, and R. Levy, "Simplified amino acid alphabets for protein fold recognition and implications for folding," *Protein Engineering*, vol. 13, no. 4, pp. 149–152, 2000.

[5] J. Li and W. Wang, "Grouping of amino acids and recognition of protein structurally conserved regions by reduced alphabets of amino acids," *Science in China Series C: Life Sciences*, vol. 50, no. 3, pp. 392–402, 2007.

[6] S. Karlin and G. Ghandour, "Multiple-alphabet amino acid sequence comparisons of the immunoglobulin kappa-chain constant domain," in *Proceedings of the National Academy of Sciences*, 1985, pp. 8597–8601.

[7] M. Sagot, A. Viari, and S. H., "Multiple comparison: a peptide matching approach," *Lecture Note in Computer Science*, vol. 937, pp. 366–385, 1995.

[8] J. Wang and W. Wang, "A computational approach to simplifying the protein folding alphabet," *Nature Structural & Molecular Biology*, vol. 6, no. 11, pp. 1033–1038, 1999.

[9] Y. Chung, Y. Kim, and H. Park, "Predicting Protein-Protein Interactions from One Feature Using SVM," in *proceedings of IEAAIE*, 2004, pp. 50–55.

[10] S. Kamtekar, J. Schiffer, H. Xiong, J. Babik, and M. Hecht, "Protein design by binary patterning of polar and nonpolar amino acids," *Science*, vol. 262, pp. 1680–1685, 1993.

[11] K. Dill, "Theory for the folding and stability of globular proteins," *Biochemistry*, vol. 24, pp. 1501–1509, 1985.

[12] S. Roy, G. Ratnaswamy, J. Boice, R. Fairman, G. McLendon, and M. Hecht, "A protein designed by binary patterning of polar and nonpolar amino acids displays native-like properties," *Jounal of the American Chemical Society*, vol. 119, no. 23, pp. 5302–5306, 1997.

[13] Wikipedia, "Amino-acids," http://en.wikipedia.org/wiki/Amino_acid.

[14] M. Park, D. Falcone, and K. Daniels, "Detection and Prediction of Alternative Splicing in Arabidopsis thaliana," *Int. J. Computational Biology and Drug Design*, vol. 1, pp. 39–58, 2008.

[15] "The Arabidopsis Information Resource," http://www.arabidopsis.org.

[16] M. Park, D. Falcone, K. Yun, and D. K., "Detection and Prediction of Alternative Splicing within Acceptor/Donor Sites in pre-mRNA of Arabidopsis thaliana," in *Proceedings of IEEE 7th International Conference on Bioinformatics & BioEngineering*, 2007, pp. 180–186.

[17] "The Institute for Genomic Research," http://www.tigr.org/.

[18] M. Pertea, X. Lin, and S. Salzberg, "GeneSplicer: a new computational method for splice site prediction," *Nucleic Acids Research*, vol. 29, pp. 1185–1190, 2001.

[19] C. Gibas and P. Jambeck, *Developing Bioinformatics Computer Skills*. O'Reilly, 2001.