Construction of Motif Vectors

The Progress Report I for the Senior Design Project: Automated Annotation of Amino Acid Sequences Using Hierarchical Motif Vectors

Student Name, Number: Mert Ekren, 200206014

Objective

The motif vectors are constructed for numerical physico-chemical properties that will represent each of amino acids. Features are extracted from online database and exposed to clustering analysis in order to obtain compact vectors.

The progress

Firstly, the project has been started from researching the physico-chemical properties that can numerically describe each 20 naturally occurring amino acids(aa). GenomeNet provides 566 indices of different physico-chemical and biological features of aa under the database "AAindex: Amino Acid Index Database". This data is parsed to the MATLAB environment with extracting 13 indices which lack of not having several numerical data for each of 20 aa.

Moving on, similarity has been sought among features in order to obtain more compact motif vectors. The correlation coefficients are implemented for the similarity parameter. It is a statistical quantification value which defines the dependence and correlation. The number ranges between minus one and one where -1.0 means the perfect negative correlation and 1.0 corresponds to the perfect positive correlation. Also, when correlation coefficient is zero, it signifies that the data is uncorrelated. In this manner, absolute value is taken into account, since the magnitude specifies the measure of correlation. A matrix of correlation coefficients among numerical properties is constructed which has the size of 533x533. It is a symmetrical square matrix where element (i,j) shows the correlation of features ith and jth. The colormapped figure of correlation coefficients is shown in Figure 1. Noting that, yellowish colors represent the high correlations whereas blueish colors signify the low correlations among the matrix.

Then, hierarchical clustering is operated to cluster similar physico-chemical features for condensed motifs. MATLAB's readily including function 'link-

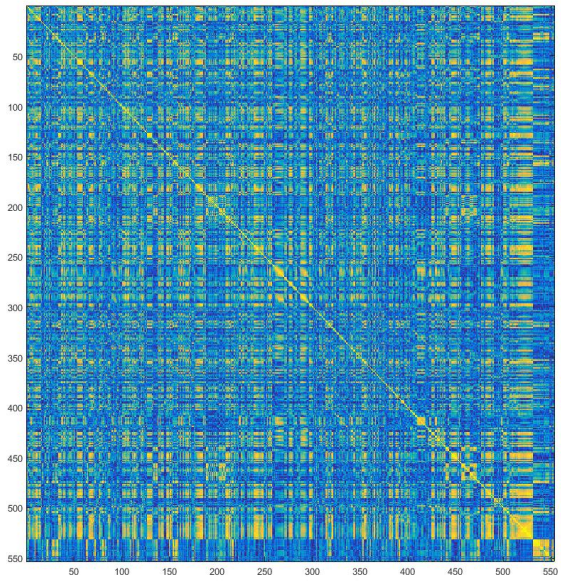

Figure 1: Colored figure of Correlation Coefficients of Features

age' is used in this manner. It is a typical example of hierarchical clustering analysis where function agglomeratively links the given distance input with respect to required distance metric. It outputs the linked points and the distance of two observations. The correlation coefficient matrix is manipulated as distance input to 'linkage' function. For the metric, three different cases are investigated which are minimum, maximum and average distance.

After linking 533 features into clusters, the similarity between clustered features has been compared with the linear diagonal line(Figure 2). The maximum difference of this comparison reveals the optimum number of clusters. This process is researched among three different distance metrics. Average distance appears as the most rational choice which suggests the number of clusters as 131. Then, for each cluster, a feature is sought to represent the its corresponding cluster. This problem is handled by similar to finding the centroid feature of each cluster. Correlation coefficients are implemented and for each feature in a cluster, the minimum correlation is found. Then, finding the maximum among this values has revealed the best possible feature for representing that cluster. This features are given in following section.

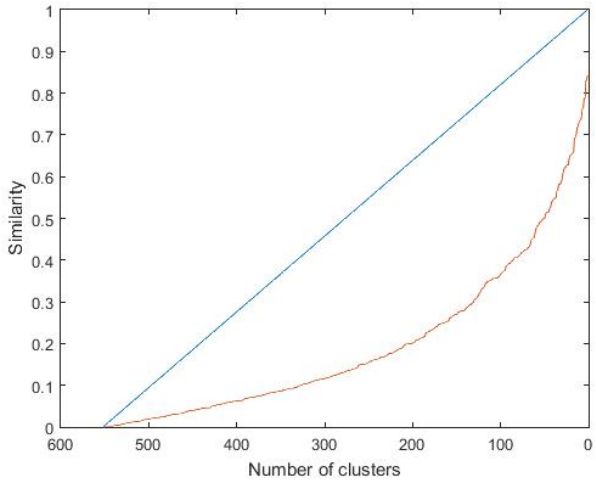

Figure 2: Number of Clusters versus similarity in comparison to linear diagonal line

Indices—Features That Represents Clusters

1 — Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga, 1982)
2 — Relative preference value at C4 (Richardson-Richardson, 1988)
3 — Relative preference value at C1 (Richardson-Richardson, 1988)
4 — Relative preference value at C2 (Richardson-Richardson, 1988)
5 — N.m.r. chemical shift of alpha-carbon (Fauchere et al., 1988)
6 — Normalized positional residue frequency at helix termini C' (Aurora-Rose,
7 — Activation Gibbs energy of unfolding, pH7.0 (Yutani et al., 1987)
8 — Unfolding Gibbs energy in water, pH9.0 (Yutani et al., 1987)
9 — Normalized frequency of extended structure (Tanaka-Scheraga, 1977)
10 — Information measure for pleated-sheet (Robson-Suzuki, 1976)
11 — A parameter of charge transfer donor capability (Charton-Charton, 1983)
12 — Entire chain compositino of amino acids in nuclear proteins (percent)
13 — Zimm-Bragg parameter sigma x 1.0E4 (Sueki et al., 1984)
14 — Weights for coil at the window position of -6 (Qian-Sejnowski, 1988)
15 — Weights for alpha-helix at the window position of -6 (Qian-Sejnowski, 1988)
16 — Hydrophobicity coefficient in RP-HPLC, C4 with 0.117 — Weights for beta-sheet at the window position of 3 (Qian-Sejnowski, 1988)
18 — Average relative fractional occurrence in AR(i) (Rackovsky-Scheraga, 1982)
19 — Weights for coil at the window position of -3 (Qian-Sejnowski, 1988)
20 — Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga, 1982)
21 — Optimized relative partition energies - method D (Miyazawa-Jernigan, 1999)
22 — Scaled side chain hydrophobicity values (Black-Mould, 1991)
23 — Composition (Grantham, 1974)
24 — Partial specific volume (Cohn-Edsall, 1943)
25 — Heat capacity (Hutchens, 1970)
26 — Distribution of amino acid residues in the alpha-helices in mesophilic
27 — Frequency of occurrence in beta-bends (Lewis et al., 1971)
28 — Information measure for loop (Robson-Suzuki, 1976)
29 — Relative preference value at C5 (Richardson-Richardson, 1988)
30 — Normalized positional residue frequency at helix termini C" (Aurora-Rose,
31 — Relative preference value at C3 (Richardson-Richardson, 1988)
32 — Alpha-helix propensity derived from designed sequences (Koehl-Levitt, 1999)
33 — Amphiphilicity index (Mitaku et al., 2002)
34 — Weighted maximum eigenvalue based on the atomic numbers
35 — Ratio of average and computed composition (Nakashima et al., 1990)
36 — pK (-COOH) (Jones, 1975)
37 — Principal property value z3 (Wold et al., 1987)
38 — Normalized composition from fungi and plant (Nakashima et al., 1990)
39 — Hydrophobicity coefficient in RP-HPLC, C18 with 0.140 — Information measure for N-terminal turn (Robson-Suzuki, 1976)
41 — Normalized positional residue frequency at helix termini N"' (Aurora-Rose,
42 — Normalized frequency of turn in all-alpha class (Palau et al., 1981)
43 — The number of atoms in the side chain labelled 1+1 (Charton-Charton, 1983)
44 — Thermodynamic beta sheet propensity (Kim-Berg, 1993)
45 — Polarity (Zimmerman et al., 1968)
46 — Principal component II (Sneath, 1966)
47 — Weights for alpha-helix at the window position of 6 (Qian-Sejnowski, 1988)
48 — Weights for coil at the window position of 5 (Qian-Sejnowski, 1988)
49 — Relative preference value at N1 (Richardson-Richardson, 1988)
50 — Normalized positional residue

frequency at helix termini N1 (Aurora-Rose,
51 — Normalized positional residue frequency at helix termini Cc (Aurora-Rose,
52 — Second smallest eigenvalue of the Laplacian matrix of the graph
53 — Weighted second smallest eigenvalue of the weighted Laplacian matrix
54 — Relative preference value at C" (Richardson-Richardson, 1988)
55 — Normalized frequency of isolated helix (Tanaka-Scheraga, 1977)
56 — Surface composition of amino acids in intracellular proteins of mesophiles
57 — Normalized frequency of alpha-helix in alpha+beta class (Palau et al., 1981)
58 — Weights for alpha-helix at the window position of -5 (Qian-Sejnowski, 1988)
59 — Free energy in alpha-helical region (Munoz-Serrano, 1994)
60 — A parameter defined from the residuals obtained from the best correlation of
61 — pK-a(RCOOH) (Fauchere et al., 1988)
62 — Alpha helix propensity of position 44 in T4 lysozyme (Blaber et al., 1993)
63 — Flexibility parameter for two rigid neighbors (Karplus-Schulz, 1985)
64 — Beta-sheet propensity derived from designed sequences (Koehl-Levitt, 1999)
65 — Free energy of solution in water, kcal/mole (Charton-Charton, 1982)
66 — Linker propensity index (Suyama-Ohara, 2003)
67 — Helix termination parameter at posision j+1 (Finkelstein et al., 1991)
68 — Net charge (Klein et al., 1984)
69 — Normalized positional residue frequency at helix termini N4'(Aurora-Rose,
70 — Normalized positional residue frequency at helix termini C4' (Aurora-Rose,
71 — Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi, 2002)
72 — Relative population of conformational state C (Vasquez et al., 1983)
73 — The Kerr-constant increments (Khanarian-Moore, 1980)
74 — Average weighted atomic number or degree based on atomic number in the graph
75 — Linker propensity from medium dataset (linker length is between six and 14
76 — Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)
77 — Weights for beta-sheet at the window position of -4 (Qian-Sejnowski, 1988)
78 — Relative preference value at N4 (Richardson-Richardson, 1988)
79 — Normalized frequency of zeta R (Maxfield-Scheraga, 1976)
80 — Normalized frequency of chain reversal S (Tanaka-Scheraga, 1977)
81 — The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)
82 — Optimized average non-bonded energy per atom (Oobatake et al., 1985)
83 — Hydrophobicity coefficient in RP-HPLC, C18 with 0.184 — Weights for coil at the window position of -5 (Qian-Sejnowski, 1988)
85 — Free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga, 1978)
86 — Average relative fractional occurrence in A0(i) (Rackovsky-Scheraga, 1982)
87 — Normalized frequency of chain reversal D (Tanaka-Scheraga, 1977)
88 — A parameter of charge transfer capability (Charton-Charton, 1983)
89 — Normalized frequency of N-terminal non helical region (Chou-Fasman, 1978b)
90 — Normalized relative frequency of double bend (Isogai et al., 1980)
91 — Principal component IV (Sneath, 1966)
92 — Weights for beta-sheet at the window position of 5 (Qian-Sejnowski, 1988)
93 — Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga, 1982)
94 — Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga, 1982)
95 — Linker propensity from small dataset (linker length is less than six
96 — Free energy change of epsilon(i) to alpha(Rh) (Wertz-Scheraga, 1978)
97 — Normalized frequency of alpha region (Maxfield-Scheraga, 1976)
98 — Linker propensity from long dataset (linker length is greater than 14
99 — Bulkiness (Zimmerman et al., 1968)
100 — Relative mutability (Jones et al., 1992)
101 — Dependence of partition coefficient on ionic strength (Zaslavsky et al.,
102 — alpha-CH chemical shifts (Bundi-Wuthrich, 1979)
103 — Surrounding hydrophobicity in beta-sheet (Ponnuswamy et al., 1980)
104 — Principal component I (Sneath, 1966)
105 — Loss of Side chain hydropathy by helix formation (Roseman, 1988)
106 — Interior composition of amino acids in nuclear proteins (percent)
107 — Relative preference value at N' (Richardson-Richardson, 1988)
108 — Relative population of conformational state A (Vasquez et al., 1983)
109 — Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)
110 — Normalized frequency of C-terminal non helical region (Chou-Fasman, 1978b)
111 — Relative preference value at C' (Richardson-Richardson, 1988)
112 — Relative preference value at N3 (Richardson-Richardson, 1988)
113 — Hydrostatic pressure asymmetry index, PAI (Di Giulio, 2005)
114 — Relative preference value at N5 (Richardson-Richardson, 1988)
115 — Correlation coefficient in regression analysis (Prabhakaran-Ponnuswamy, 1982)
116 — Alpha-helix indices for beta-proteins (Geisow-Roberts, 1980)
117 — Weights for beta-sheet at the window position of -3 (Qian-Sejnowski, 1988)
118 — Normalized positional residue frequency at helix termini N'(Aurora-Rose,
119 — Hydrophobicity (Zimmerman et al., 1968)
120 — Melting point (Fasman, 1976)
121 — Information measure for C-terminal turn (Robson-Suzuki, 1976)
122 — Information measure for extended without H-bond (Robson-Suzuki, 1976)
123 — Normalized frequency of zeta R (Tanaka-Scheraga, 1977)
124 — Hydropathy scale based on self-information values in the two-state model (50125 — Normalized positional residue frequency at helix termini N" (Aurora-Rose,
126 — Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
127 — Weights for coil at the window position of 6 (Qian-Sejnowski, 1988)
128 — Weighted minimum eigenvalue based on the atomic numbers
129 — Intercept in regression analysis (Prabhakaran-Ponnuswamy, 1982)
130 — Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich, 1979)
131 — Electron-ion interaction potential values (Cosic, 1994)