Biophysics, Vol. 48, Suppl. 1, 2003, pp. S110–S115.

Profile Clusters Derived from BLOCKS Suggest a Simple Model of Column Evolution in Multiple Alignments of Protein Families

I. V. Merkeev and A. A. Mironov

State Scientific Center GosNIIGenetica, Moscow, 113545 Russia; E-mail: ivmerkeev@rambler.ru

Received November 10, 2003; in final form, December 20, 2003

Abstract—BLOSUM and PAM series of protein substitution matrices are popular tools for scoring protein pairwise and multiple alignments. For protein multiple alignments there exists another representation of an evolving column, based on a set of predefined frequency profile patterns. For conserved sites, these profile patterns represent stationary points in a 20-dimensional profile space. There are 20 such patterns. All of them were derived from the BLOCKS database by applying a special clusterization procedure to frequency profiles obtained from BLOCKS alignment columns. To understand the nature of these clusters, random protein sequences were generated where all columns were obtained from a single amino acid type by applying transition probabilities to it, and the same clusterization procedure was applied to the generated frequency profiles. Similar twenty clusters were obtained. This means that for conservative columns, all amino acids in that column are derived from a single ancestor amino acid by a substitution random process with standard transition probabilities. For non-conservative columns there are, generally, no regularities in the amino acid types present therein. Based on the COG database, a formula was obtained to distinguish between functionally important and unimportant columns. If the odds ratio of likelihoods of the most probable ancestral amino acid to the third most probable ancestral amino acid exceeds a critical value, then this column is predicted to be conservative and it is the result of evolution of the ancestral amino acid. Otherwise, this column is considered to be a "garbage" column. When building a consensus sequence from a multiple alignment, we can represent this column as a "garbage" symbol having zero value with any amino acid in a substitution matrix.

Key words: SNP, profile clusters, column evolution, amino acid substitution matrix, protein evolution model

INTRODUCTION

Profiles introduced by Gribskov and co-workers [1] have proved to be a valuable tool for finding weak homologies between distant proteins belonging to one family or superfamily and for improving sensitivity of database searches [2,3]. Two views have been established on the composition of profiles: continuous and discrete. The continuous model of profiles imposes no restriction on amino acids and their counts present in an alignment column, while the discrete model assumes that only certain amino acids can be found at a certain position in the multiple alignment of a protein family.

The efforts of several workers were directed to improve profiles by sequence weighting [4–8] and by introducing pseudocounts [9–12]. As pointed out by Tatusov and co-workers [2], the most efficient method in improving the quality of profiles by adding pseudocounts is the Dirichlet mixture. This mixture is a linear combination of nine Dirichlet distributions. Although no restrictions are imposed on frequency profiles in this model, nine Dirichlet distributions can be considered as discrete condensation points in a 20-dimensional space.

The discrete view on profiles is exemplified by the PROSITE database [13], which states what kind of amino acid can be present in a single position of a protein signature.

In this work, we will present strong evidence that all 20 profiles can be regarded as a result of evolution of just one ancestral amino acid. If a multiple protein alignment has enough sequences, we can

Table 1. Average profiles of frequency profile clusters obtained from BLOCKS. Dominating amino acids are in boldface

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	**0.486**	0.022	0.014	0.023	0.015	0.062	0.008	0.030	0.023	0.038	0.015	0.016	0.016	0.017	0.019	0.081	0.041	0.059	0.003	0.011
2	0.039	**0.702**	0.010	0.010	0.014	0.019	0.007	0.014	0.009	0.024	0.008	0.014	0.010	0.008	0.011	0.033	0.025	0.030	0.004	0.010
3	0.032	0.004	**0.522**	0.081	0.008	0.030	0.018	0.011	0.035	0.016	0.006	0.069	0.017	0.026	0.022	0.048	0.028	0.015	0.003	0.009
4	0.055	0.004	0.086	**0.427**	0.009	0.023	0.015	0.017	0.066	0.026	0.009	0.032	0.020	0.057	0.039	0.045	0.033	0.024	0.003	0.011
5	0.027	0.008	0.008	0.011	**0.528**	0.012	0.012	0.046	0.011	0.092	0.023	0.009	0.009	0.009	0.013	0.019	0.020	0.040	0.021	0.080
6	0.053	0.007	0.025	0.020	0.010	**0.658**	0.009	0.011	0.024	0.015	0.006	0.033	0.013	0.013	0.018	0.043	0.019	0.016	0.003	0.007
7	0.025	0.005	0.022	0.021	0.027	0.017	**0.581**	0.014	0.025	0.024	0.007	0.040	0.011	0.035	0.026	0.028	0.018	0.016	0.008	0.048
8	0.033	0.009	0.007	0.010	0.039	0.011	0.006	**0.437**	0.012	0.138	0.031	0.009	0.008	0.008	0.011	0.016	0.026	0.169	0.005	0.016
9	0.045	0.005	0.028	0.057	0.010	0.023	0.019	0.022	**0.401**	0.035	0.011	0.038	0.017	0.051	0.113	0.044	0.038	0.027	0.003	0.012
10	0.036	0.010	0.008	0.014	0.052	0.013	0.008	0.103	0.015	**0.500**	0.049	0.009	0.009	0.013	0.016	0.019	0.023	0.077	0.007	0.019
11	0.045	0.010	0.008	0.015	0.039	0.013	0.008	0.064	0.013	0.139	**0.450**	0.014	0.008	0.026	0.015	0.024	0.029	0.060	0.006	0.016
12	0.033	0.007	0.062	0.033	0.011	0.043	0.029	0.015	0.042	0.021	0.008	**0.476**	0.014	0.030	0.033	0.066	0.042	0.017	0.003	0.015
13	0.055	0.005	0.027	0.034	0.013	0.025	0.011	0.019	0.036	0.030	0.007	0.018	**0.559**	0.020	0.025	0.045	0.030	0.029	0.004	0.010
14	0.041	0.004	0.029	0.069	0.012	0.020	0.024	0.020	0.065	0.038	0.019	0.036	0.012	**0.435**	0.054	0.042	0.034	0.028	0.004	0.013
15	0.035	0.006	0.018	0.032	0.012	0.021	0.022	0.019	0.094	0.033	0.011	0.028	0.013	0.038	**0.509**	0.035	0.029	0.023	0.006	0.015
16	0.087	0.014	0.034	0.029	0.014	0.051	0.015	0.017	0.030	0.024	0.011	0.044	0.020	0.023	0.023	**0.436**	0.087	0.025	0.004	0.014
17	0.053	0.013	0.022	0.024	0.014	0.022	0.012	0.030	0.030	0.035	0.014	0.032	0.013	0.022	0.024	0.107	**0.470**	0.050	0.003	0.011
18	0.061	0.016	0.008	0.015	0.026	0.013	0.007	0.145	0.017	0.088	0.021	0.009	0.011	0.010	0.014	0.023	0.045	**0.453**	0.004	0.015
19	0.025	0.006	0.011	0.014	0.067	0.015	0.014	0.026	0.015	0.046	0.016	0.012	0.011	0.010	0.018	0.022	0.016	0.024	**0.579**	0.052
20	0.024	0.007	0.010	0.015	0.120	0.014	0.033	0.026	0.017	0.045	0.013	0.017	0.010	0.012	0.018	0.022	0.019	0.031	0.026	**0.521**

assign any column to a profile cluster with a high degree of confidence.

METHODS

Computing Frequency Profile Clusters from BLOCKS

To find profile clusters, we used the BLOCKS database Version 12.0 [14]. Each block from BLOCKS was processed to give a multiple alignment of sequences in the following way. The first sequence was extracted from the block. Assume that we processed a certain number of sequences in a block. If the next sequence to be processed is at ≤65% identity with the previous sequences that form the growing multiple alignment. Thus all sequences in the resulting multiple alignment are at ≤65% identity with each other. If we convert blocks to multiple alignments without such filtering many similar sequences can distort frequency profiles. Only alignments having at least 15 sequences were considered for further processing to have enough statistical material. Totally, 39 253 alignment columns were obtained. The average number of sequences in a multiple alignment was 32. The average protein identity in a multiple alignment was 27%.

Then each alignment column was converted to the frequency profile using a simple formula:

$$f_i = \frac{N_i}{N}$$

$$\vec{f} = (f_1, f_2, ..., f_{20}),$$

where N_i is the number of times amino acid i occurs in the column, N is the number of sequences in the multiple alignment.

Table 2. Average profiles of frequency profile clusters obtained from randomly generated columns. Dominating amino acids are in boldface

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	**0.480**	0.000	0.027	0.027	0.013	0.055	0.014	0.027	0.028	0.028	0.014	0.027	0.014	0.027	0.014	0.082	0.042	0.067	0.000	0.014
2	0.046	**0.663**	0.000	0.015	0.015	0.015	0.015	0.015	0.015	0.046	0.000	0.016	0.015	0.000	0.015	0.031	0.031	0.031	0.000	0.016
3	0.043	0.000	**0.560**	0.084	0.000	0.043	0.000	0.013	0.027	0.028	0.000	0.059	0.015	0.015	0.013	0.042	0.027	0.015	0.000	0.014
4	0.042	0.014	0.068	**0.502**	0.000	0.014	0.028	0.000	0.069	0.029	0.013	0.028	0.014	0.055	0.027	0.040	0.015	0.027	0.000	0.015
5	0.029	0.015	0.000	0.015	**0.543**	0.014	0.000	0.042	0.015	0.100	0.015	0.015	0.015	0.000	0.014	0.015	0.027	0.028	0.014	0.085
6	0.060	0.014	0.015	0.016	0.015	**0.653**	0.000	0.015	0.031	0.016	0.000	0.029	0.015	0.015	0.015	0.045	0.015	0.015	0.000	0.016
7	0.028	0.014	0.015	0.043	0.014	0.028	**0.546**	0.000	0.029	0.028	0.014	0.042	0.013	0.044	0.029	0.029	0.014	0.014	0.000	0.056
8	0.026	0.013	0.013	0.000	0.026	0.013	0.014	**0.434**	0.013	0.129	0.027	0.013	0.013	0.012	0.000	0.026	0.014	0.186	0.000	0.027
9	0.039	0.014	0.027	0.067	0.014	0.027	0.000	0.027	**0.455**	0.014	0.026	0.027	0.026	0.039	0.092	0.039	0.026	0.013	0.014	0.014
10	0.028	0.013	0.013	0.014	0.056	0.013	0.000	0.097	0.014	**0.512**	0.057	0.000	0.014	0.014	0.014	0.014	0.028	0.069	0.000	0.028
11	0.050	0.013	0.013	0.012	0.024	0.013	0.025	0.063	0.050	0.172	**0.364**	0.013	0.013	0.024	0.013	0.025	0.024	0.062	0.000	0.025
12	0.042	0.000	0.069	0.027	0.013	0.055	0.013	0.027	0.042	0.014	0.013	**0.493**	0.014	0.028	0.028	0.054	0.041	0.013	0.000	0.013
13	0.046	0.015	0.016	0.029	0.000	0.030	0.015	0.016	0.031	0.030	0.000	0.016	**0.635**	0.015	0.015	0.031	0.029	0.016	0.000	0.016
14	0.054	0.013	0.014	0.092	0.013	0.014	0.027	0.013	0.065	0.040	0.013	0.027	0.013	**0.433**	0.053	0.053	0.026	0.013	0.000	0.025
15	0.028	0.014	0.014	0.030	0.014	0.014	0.027	0.000	0.115	0.027	0.014	0.015	0.014	0.043	**0.530**	0.028	0.014	0.029	0.000	0.028
16	0.105	0.014	0.027	0.039	0.013	0.051	0.014	0.012	0.040	0.026	0.000	0.040	0.026	0.027	0.013	**0.427**	0.089	0.012	0.014	0.013
17	0.066	0.000	0.027	0.026	0.012	0.027	0.000	0.039	0.027	0.040	0.013	0.041	0.027	0.013	0.013	0.106	**0.445**	0.053	0.013	0.013
18	0.065	0.014	0.013	0.014	0.026	0.013	0.000	0.155	0.014	0.095	0.026	0.000	0.013	0.013	0.013	0.013	0.052	**0.434**	0.000	0.026
19	0.030	0.000	0.016	0.015	0.031	0.031	0.016	0.015	0.015	0.015	0.016	0.014	0.000	0.016	0.015	0.014	0.014	0.015	**0.652**	0.060
20	0.028	0.000	0.014	0.014	0.112	0.013	0.029	0.027	0.028	0.057	0.013	0.014	0.014	0.014	0.014	0.027	0.013	0.041	0.013	**0.515**

To find the similarity score between frequency profiles, we used the Pearson correlation coefficient r:

$$r(\vec{f}', \vec{f}'') = \frac{\sum_{i=1}^{20} (f_i' - \bar{f})(f_i'' - \bar{f}'')}{\sqrt{\sum_{i=1}^{20} (f_i' - \bar{f})^2 \sum_{i=1}^{20} (f_i'' - \bar{f}'')^2}}.$$

We have chosen the Pearson correlation coefficient as the comparison measure between profiles largely because of the results obtained by Pietrokovski [15], who studied how to find homologies between blocks in the BLOCKS database and found that of four different similarity measures between profiles the Pearson correlation coefficient gave the best results. For each frequency profile, similar profiles were found that had therewith a Pearson correlation coefficient of at least 0.7. Profiles already included in profile groups were not considered as seeds of new groups. For each of these groups, an average profile was calculated. Then an iterative procedure was launched. At each step each profile was assigned to the cluster with the nearest average. The averages were recomputed and very similar clusters were merged. The procedure came to termination when the number of clusters became stable.

Table 1 shows 20 average profiles corresponding to each frequency cluster. Each cluster is dominated by a single amino acid.

Random Test

To understand the nature of profile clusters, a random test was performed, generating profiles from random protein sequences. Parameters of the generation process were chosen to closely mimic the parameters of the previous procedure. 400 random protein sequences each having a length of 100 amino acids were generated with standard background frequencies of amino acids. This gave us 40 000 alignment columns, which is close to the number of alignment columns used in the previous procedure (39 253). Each generated sequence was considered as an ancestral sequence, and the process of protein evolution in each column of the generated sequence was modeled by random generation of 32 amino acids with probabilities determined by the amino acid substitution matrix BLOSUM62. This number corresponds to the average number of sequences in a multiple alignment obtained in the previous procedure. The average protein identity in generated protein sequences was 28.7%, which is close to the value obtained in the previous procedure (27%).

The same clusterization procedure was applied to randomly generated profiles, and similar 20 clusters were obtained (Table 2). Table 3 shows Pearson correlation coefficients between corresponding profile averages from both groups of clusters.

Table 3. Pearson correlation coefficients between average profiles obtained from BLOCKS and average profiles obtained from random protein sequences

Cluster index	Pearson correlation coefficient
1	0.997
2	0.998
3	0.997
4	0.994
5	0.998
6	0.999
7	0.997
8	0.996
9	0.990
10	0.998
11	0.982
12	0.997
13	0.998
14	0.994
15	0.996
16	0.996
17	0.996
18	0.997
19	0.995
20	0.998

Test whether an Alignment Column Belongs to a Profile Cluster

In multiple alignments it is very important to know whether a column belongs to the cluster or not. If we know that a column belongs to a particular cluster, it is possible to replace it with a single symbol corresponding to the ancestral amino acid. If the column does not belong to any cluster, it is called a "garbage" column. This means that there are no regularities in amino acid types and counts present in that column.

Intuitively, our decision to assign a particular alignment column to the profile cluster will depend on the number of sequences in the multiple alignment, because the more there are sequences the more information we have to make a decision. In the extreme case, when we have only one sequence, it will automatically belong to its cluster, since information is too scarce to make a decision.

We used the ratio of the likelihoods of the most probable ancestral amino acid to the likelihood of the third most probable amino acid as our decision rule (p_1/p_3). Likelihoods for p_α were calculated using the following formula:

$$p_\alpha = f_\alpha \cdot \prod_{\beta=1}^{20} q_{\alpha\beta}^{n_\beta}$$

where f_α is the background frequency of amino acid α, $q_{\alpha\beta}$ is the transition probability from amino acid β to the amino acid β, n_β is the number of times amino acid β occurs in the column, p_α is the likelihood that amino acid α was the ancestral amino acid. Likelihoods for all 20 amino acids were then sorted in decreasing order: $p_1 \geq p_2 \geq \ldots \geq p_{20}$.

If the ratio p_1/p_3 exceeds a critical value, then the column will belong to the cluster, otherwise it will be declared as a "garbage" symbol. To find out how this critical value depends on N, the number of sequences in the multiple alignment, we used the COG database [16]. For different N we performed the following experiments. N sequences were extracted randomly from each COG, and then another N sequences were extracted randomly from the sequences that were left

in this COG after the first extraction. As a result of this procedure, each COG gave us two multiple alignments each containing N sequences. Then different critical values were tested. In each such test, multiple alignments were converted to consensus sequences following the decision rule p_1/p_3. Since each COG gave two consensus sequences, all consensus sequences were partitioned into two groups, and then TOG-BLAST [17] was launched to find bidirectional best hits (BBHs) between these two groups. Then the critical value was found corresponding to the maximum number of BBHs.

Basing on the results of such computational experiments (data not shown), the following formula was derived:

$$\frac{p_1}{p_3} > 0.8 + 20\left(\frac{N}{10}\right)^2.$$

RESULTS AND DISCUSSION

Profile clusters obtained from BLOCKS demonstrate that, despite the seeming infiniteness of profile space, it is possible to find condensation points for functionally important columns and to indicate their evolutionary origin. To the best of our knowledge, this is the first attempt to clusterize BLOCKS, therefore much is remained to investigate. Many intriguing questions arise. Are there any subtypes of the main types that will give finer classification of condensation points? What is the probability distribution of profiles around their average points? Do such clusters exist in other alignments of protein families? If multiple alignments naturally tend to form clusters in their columns, the maximization of the sum of pairs as the usual optimization goal for multiple alignments can be substituted by the maximization of values p_1/p_3 summed over all alignments columns. This can potentially result in developing new algorithms for multiple alignment of sequences.

One very useful application of profile clusters, which was immediately used after their discovery, is building supergenomes from complete proteome complements [17]. One of the tasks in this process is to run a homology search program between multiple alignments to find BBHs between them. Multiple alignments are converted into consensus sequences using the concept of profile clusters. Positive results obtained in building supergenomes prove the usefulness of profile clusters.

ACKNOWLEDGMENTS

We are grateful to S. Sunyaev for discussion.

This study was partially supported by grants from HHMI (55000309) and LICR (CRDF RBO-1268).

REFERENCES

1. Gribskov, M., McLachlan, A.D., and Eisenberg, D., Profile analysis: Detection of distantly related proteins, *Proc. Natl. Acad. Sci. USA*, 1987, vol. 84, pp. 4355–4358.
2. Tatusov, R.L., Altschul, S.F., and Koonin, E. V., Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignments blocks, *Proc. Natl. Acad. Sci. USA*, 1994, vol. 91, pp. 12091–12095.
3. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acid Research*, 1997, vol. 25, pp. 3389–3402.
4. Altschul, S.F., Carroll, R.J., and Lipman, D.J., Weights for data related by a tree, *J. Mol. Biol.*, 1989, vol. 207, pp. 647–653.
5. Sibbald, P.R. and Argos, P., Weighting aligned protein or nucleic acid sequences to correct for unequal representation, *J. Mol. Biol.*, 1990, vol. 216, pp. 813–818.
6. Vingron, M. and Sibbald, P.R., Weighting in sequence space: A comparison of methods in terms of generalized sequences, *Proc. Natl. Acad. Sci. USA*, 1993, vol. 90, pp. 8777–8781.
7. Thompson, J.D., Higgins, D.G., and Gibson, T.J., Improved sensitivity of profile searches through the use of sequence weights and gap excision, *CABIOS*, 1994, vol. 10, pp. 19–29.
8. Henikoff, S. and Henikoff, J.G., Position-based sequence weights, *J. Mol. Biol.*, 1994, vol. 243, pp. 574–578.
9. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C., Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment, *Science*, 1993, vol. 262, pp. 208–214.
10. Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., and Haussler, D., Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology, *CABIOS*, 1996, vol. 12, pp. 327–345.

11. Henikoff, J.G. and Henikoff, S., Using substitution probabilities to improve position-specific scoring matrices, CABIOS, 1996, vol. 12, pp. 135–143.

12. Sunyaev, S.R., Eisenhaber, F., Rodchenkov, I.V., Eisenhaber, B., Tumanyan, V.G., and Kuznetsov, E.N. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations, *Protein Engineering*, 1999, vol. 12, pp. 387–394.

13. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K., and Bairoch, A., The PROSITE database, its status in 2002, *Nucleic Acid Research*, 2002, vol. 30, pp. 235–238.

14. Henikoff, S. and Henikoff, J.G., Automated assembly of protein blocks for database searching, *Nucleic Acids Research*, 1991, vol. 19, pp. 6565–6572.

15. Pietrokovski, S., Searching databases of conserved sequence regions by aligning protein multiple-alignments, *Nucleic Acid Research*, 1996, vol. 24, pp. 3836–3845.

16. Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V., The COG database: new developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acid Research*, 2001, vol. 29, pp. 22–28.

17. Merkeev, I.V., Novichkov, P.S., and Mironov, A.A., TOGs vs. COGs: A database of supergenomes built from complete proteome complements, 2003, Article in press.