Стукалов А. С.

МГУ им. М. В. Ломоносова, ast@pochta.ru

ОЦЕНКА КАЧЕСТВА КЛАСТЕРИЗАЦИИ ВЫРОВНЕННЫХ СТРОК

Вопрос нахождения кластеров в наборе строк одинаковой длины является одним из актуальных вопросов биоинформатики. Вводится векторный критерий, позволяющий оценивать качество таких кластеризаций, рассматриваются способы его свертки. Данный критерий может быть использован для построения методов кластеризации.

1. Введение

Кластеризация данных [1], т. е. разбиение набора данных на подмножества, каждое из которых обладает некоторым специфическим для его элементов свойством, находит много применений в биоинформатике, экономике, распознавании образов и т. д. [2]

В частности, в биоинформатике кластеризация используется для выделения семейств гомологичных белков [3, 5, 6]. Обычно входные данные — это *множественное выравнивание* (*multiple sequence alignment*, *MSA*), набор выровненных друг относительно друга белковых последовательностей. MSA представляет собой матрицу, каждая строка которой есть последовательность аминокислотных остатков некоторого белка. В эту строку вставлены специальные *символы-пробелы* таким образом, чтобы каждая колонка матрицы содержала преимущественно одинаковые или похожие остатки.

Одна из особенностей матрицы MSA состоит в том, что кроме *консервативных* колонок, состоящих из похожих аминокислотных остатков, в ней есть колонки, содержащие остатки существенно различной природы. Наличие таких колонок объясняется двумя причинами. Это, во-первых, функциональные и структурные различия белков, образующих матрицу. Во-вторых, *молчащие мутации*, которые не приводят к изменению структуры или функций белка.

Идеальный метод кластеризации белков должен уметь отличать колонки MSA, которые коррелированы со свойствами кластеров (*специфические колонки*) от неинформативных колонок, в которых преобладают

молчащие мутации. Специфические колонки должны быть ключевыми признаками, по которым ведется кластеризация: последовательности, содержащие на специфических позициях одинаковые остатки, должны попадать в один кластер. Кластеры, полученные таким способом, будут однозначно характеризоваться набором остатков на специфических позициях.

Можно задаться вопросом оценки качества кластеризации, введения некоторой меры, которая бы оценивала свойства, требуемые от хорошей кластеризации. Подобная мера могла бы быть использована в методе кластеризации, основанном на оптимизации некоторого функционала на множестве кластеризаций. Одним из примеров такого подхода является метод оптимизации функции контраста энтропии [5].

В данной статье предлагается двухкомпонентный векторный критерий оценки качества кластеризации. Его первый компонент оценивает *неоднородность строк*, составляющих кластер. Второй компонент оценивает *неоднозначность классифицирования*, возникающую из-за похожести строк разных кластеров друг на друга.

2. Постановка задачи

Пусть $A_1, A_2, \ldots, A_M$ — набор слов длины N в алфавите $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_L\}$. Эти слова можно считать строками матрицы $A \in \Lambda^{M \times N}$. Пусть $\widetilde{\mathcal{A}}$ — разбиение множества строк, т. е. $\widetilde{\mathcal{A}} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m\}$, где

$$\mathcal{A}_i = \{A_{i_1}, A_{i_2}, \ldots, A_{i_{n_i}}\} \neq \varnothing, \qquad \bigsqcup_{\mathcal{A}_i \in \widetilde{\mathcal{A}}} \mathcal{A}_i = \{A_1, A_2, \ldots A_M\}. \tag{1}$$

Пусть $\mathfrak{A}$ — множество всех таких разбиений $\widetilde{\mathcal{A}}$.

Предположим, существует мера *качества кластеризации*, или стоимости кластеризации, $q\colon \mathfrak{A} \to \mathbb{R}$. Задача состоит в нахождении оптимального разбиения $\widetilde{\mathcal{A}}*$:

$$q(\widetilde{\mathcal{A}}*) \leqslant q(\widetilde{\mathcal{A}}) \quad \forall\, \widetilde{\mathcal{A}} \in \mathfrak{A}. \tag{2}$$

В следующих пунктах будет рассмотрен один из возможных способов задания функции $q(\widetilde{\mathcal{A}})$.

3. Мера неоднородности состава

Пусть на множестве пар символов алфавита Λ определена функция $\rho\colon \Lambda \times \Lambda \to \mathbb{R}^+$, которая имеет смысл расстояния между двумя символами: чем больше различие между α и $\beta \in \Lambda$, тем больше $\rho(\alpha, \beta)$ (однако выполнение тождества $\rho(\alpha, \alpha) = 0$ не требуется). Простейший пример такой ρ — это *дискретная метрика* на Λ:

$$\rho(\alpha, \beta) = \begin{cases} 0, & \alpha = \beta; \\ 1, & \alpha \neq \beta. \end{cases} \tag{3}$$

Для задания расстояния в кластеризациях MSA могут быть использованы также *матрицы замен*, такие как BLOSUM или PAM [4], отражающие наблюдаемые частоты мутаций аминокислотных остатков.

Рассмотрим символы на j-й позиции в подмножестве строк $\mathcal{A} = \{A_{i_1}, A_{i_2}, \ldots, A_{i_{n_\mathcal{A}}}\}$. Пусть $n_\mathcal{A}(\lambda, j)$ — количество появлений $\lambda \in \Lambda$ в j-й колонке $\mathcal{A}$.

Определение 1. *Неоднородность состава* множества строк $\mathcal{A}$ в j-й колонке есть

$$d(\mathcal{A}, j) = \frac{1}{n_\mathcal{A}} \sum_{\alpha,\beta\in\Lambda} \rho(\alpha, \beta) n_\mathcal{A}(\alpha, j) n_\mathcal{A}(\beta, j). \tag{4}$$

Каждой паре $(\mathcal{A}, j)$ мы можем сопоставить функцию вероятности

$$p_{\mathcal{A},j}(\lambda) = \frac{1}{n_\mathcal{A}} n_\mathcal{A}(\lambda, j),$$

тогда $d(\mathcal{A}, j)$ можно переписать так:

$$d(\mathcal{A}, j) = n_\mathcal{A}\, \mathbb{E}\, \rho(\Lambda_1, \Lambda_2), \tag{5}$$

где Λ_1 и Λ_2 — две независимые случайные величины, принимающие значения из Λ, с функцией вероятности $p_{\mathcal{A},j}$.

Рассмотрим случай, когда ρ — это *дискретная метрика* на Λ, тогда

$$\mathbb{E}\, \rho(\Lambda_1, \Lambda_2) = p(\Lambda_1 \neq \Lambda_2).$$

Минимум $d(\mathcal{A}, j) = 0$ достигается, когда в каждой строке из $\mathcal{A}$ j-м символом является некоторый $\lambda \in \Lambda$. Напротив, если в j-й колонке $\mathcal{A}$ каждый символ встречается единожды, неоднородность состава достигает максимального значения $d(\mathcal{A}, j) = n_\mathcal{A} - 1$.

Пусть $\widetilde{\mathcal{A}} = \{\mathcal{A}_1, \mathcal{A}_2 \ldots, \mathcal{A}_{n_{\widetilde{\mathcal{A}}}}\}$ — попарно непересекающиеся подмножества строк A. Мы расширяем определение меры неоднородности строк и полагаем

$$d(\widetilde{\mathcal{A}}, j) = \sum_{i=1}^{n_{\widetilde{\mathcal{A}}}} d(\mathcal{A}_i, j). \tag{6}$$

Интересно установить, как разделение и объединение кластеров влияют на эту меру, т. е. сравнить $d(\mathcal{A}, j) + d(\mathcal{B}, j)$ и $d(\mathcal{A} \cup \mathcal{B}, j)$:

$$\begin{aligned} d(\mathcal{A}) + d(\mathcal{B}) - d(\mathcal{A} \cup \mathcal{B}) = \\ = \sum_{\alpha,\beta\in\Lambda} \rho(\alpha, \beta) \frac{n_\mathcal{A}(\alpha) n_\mathcal{A}(\beta)}{n_\mathcal{A}} + \sum_{\alpha,\beta\in\Lambda} \rho(\alpha, \beta) \frac{n_\mathcal{B}(\alpha) n_\mathcal{B}(\beta)}{n_\mathcal{B}} - \\ - \sum_{\alpha,\beta\in\Lambda} \rho(\alpha, \beta) \frac{n_{\mathcal{A}\cup\mathcal{B}}(\alpha) n_{\mathcal{A}\cup\mathcal{B}}(\beta)}{n_{\mathcal{A}\cup\mathcal{B}}} = \end{aligned}$$

$$= \sum_{\alpha,\beta\in\Lambda} \rho(\alpha,\beta)\left(\frac{n_{\mathcal{A}}(\alpha)n_{\mathcal{A}}(\beta)}{n_{\mathcal{A}}} + \frac{n_{\mathcal{B}}(\alpha)n_{\mathcal{B}}(\beta)}{n_{\mathcal{B}}} - \right.$$

$$\left. - \frac{(n_{\mathcal{A}}(\alpha)+n_{\mathcal{B}}(\alpha))(n_{\mathcal{A}}(\beta)+n_{\mathcal{B}}(\beta))}{n_{\mathcal{A}}+n_{\mathcal{B}}}\right) =$$

$$= \sum_{\alpha,\beta\in\Lambda} \rho(\alpha,\beta)\frac{(n_{\mathcal{A}}(\alpha)n_{\mathcal{B}} - n_{\mathcal{B}}(\alpha)n_{\mathcal{A}})(n_{\mathcal{A}}(\beta)n_{\mathcal{B}} - n_{\mathcal{B}}(\beta)n_{\mathcal{A}})}{n_{\mathcal{A}}n_{\mathcal{B}}(n_{\mathcal{A}}+n_{\mathcal{B}})}.$$

Если $R \in \mathbb{R}^{L\times L}$ — матрица расстояний: $R_{i,j} = \rho(\lambda_i,\lambda_j)$, и компонентами вектора $t_{\mathcal{A},\mathcal{B}} \in \mathbb{R}^L$ являются числа $t_{\mathcal{A},\mathcal{B},i} = n_{\mathcal{A}}(\lambda_i)n_{\mathcal{B}} - n_{\mathcal{B}}(\lambda_i)n_{\mathcal{A}}$, то

$$d(\mathcal{A}) + d(\mathcal{B}) - d(\mathcal{A}\cup\mathcal{B}) = \frac{1}{n_{\mathcal{A}}n_{\mathcal{B}}(n_{\mathcal{A}}+n_{\mathcal{B}})}\langle Rt_{\mathcal{A},\mathcal{B}}, t_{\mathcal{A},\mathcal{B}}\rangle. \tag{7}$$

Заметим, что для любых $\mathcal{A}$ и $\mathcal{B}$

$$\sum_{i=1}^{L} t_{\mathcal{A},\mathcal{B},i} = \sum_{i=1}^{L} n_{\mathcal{A}}(\lambda_i)n_{\mathcal{B}} - \sum_{i=1}^{L} n_{\mathcal{B}}(\lambda_i)n_{\mathcal{A}} = n_{\mathcal{A}}n_{\mathcal{B}} - n_{\mathcal{B}}n_{\mathcal{A}} = 0. \tag{8}$$

Этот факт будет использован для доказательства следующего утверждения:

Утверждение 1. *Если* $\rho(\alpha,\beta)$ *— дискретная метрика* (3), *тогда*

$$d(\mathcal{A}) + d(\mathcal{B}) \leqslant d(\mathcal{A}\cup\mathcal{B}) \quad \forall\, \mathcal{A},\mathcal{B},$$

$d(\mathcal{A}) + d(\mathcal{B}) = d(\mathcal{A}\cup\mathcal{B})$ *тогда и только тогда, когда*

$$n_{\mathcal{A}}(\lambda_i) = n_{\mathcal{B}}(\lambda_i) \quad \forall\, \lambda_i \in \Lambda.$$

Доказательство. Легко видеть, что матрица расстояний R для дискретной метрики имеет два различных собственных значения: $(L-1)$ с собственным вектором $(-1,-1,\ldots,-1)$ и -1 степени $L-1$ с собственным подпространством

$$T \subset \mathbb{R}^L : \sum_{i=1}^{L} t_i = 0 \quad \forall\, t \in T.$$

Таким образом, для любых $\mathcal{A}$ и $\mathcal{B}$, соответствующий им вектор $t_{\mathcal{A},\mathcal{B}}$ всегда будет содержаться в T и

$$d(\mathcal{A}) + d(\mathcal{B}) - d(\mathcal{A}\cup\mathcal{B}) = \frac{1}{n_{\mathcal{A}}n_{\mathcal{B}}(n_{\mathcal{A}}+n_{\mathcal{B}})}\langle Rt_{\mathcal{A},\mathcal{B}}, t_{\mathcal{A},\mathcal{B}}\rangle =$$

$$= \frac{-1}{n_{\mathcal{A}}n_{\mathcal{B}}(n_{\mathcal{A}}+n_{\mathcal{B}})}\|t_{\mathcal{A},\mathcal{B}}\|^2. \quad \square$$

Таким образом, объединение кластеров увеличивает неоднородность состава в случае дискретной метрики.

4. Мера неоднозначности классифицирования

Определение 2. Для данного набора $\widetilde{\mathcal{A}}$ непересекающихся подмножеств строк A и столбца j, мера *неоднозначности классифицирования*, $a(\widetilde{\mathcal{A}}, j)$, есть

$$a(\widetilde{\mathcal{A}}, j) = \sum_{\lambda\in\Lambda}\sum_{\mathcal{A}\in\widetilde{\mathcal{A}}} \frac{n_{\mathcal{A}}(\lambda, j)\big(n_{\widetilde{\mathcal{A}}}(\lambda, j) - n_{\mathcal{A}}(\lambda, j)\big)}{n_{\widetilde{\mathcal{A}}}(\lambda, j)}. \tag{9}$$

Мера неоднозначности классифицирования $a(\widetilde{\mathcal{A}}, j)$ может быть проинтерпретирована в терминах теории вероятности следующим образом. Рассмотрим двух игроков. Первый случайным образом выбирает строку s из $\widetilde{\mathcal{A}}$ и сообщает символ λ в j-й позиции этой строки другому игроку, при этом храня номер кластера $\widetilde{\mathcal{A}}_1$, из которого эта строка была взята, в секрете. Второй игрок, используя информацию о λ, пытается угадать этот кластер. Вероятность того, что второй игрок даст неверный ответ будет равна

$$p(\widetilde{\mathcal{A}}_1 \neq \widetilde{\mathcal{A}}_2) = \frac{a(\widetilde{\mathcal{A}}, j)}{N_{\widetilde{\mathcal{A}}}}, \tag{10}$$

где $\widetilde{\mathcal{A}}_2$ — предположение 2-го игрока и $N_{\widetilde{\mathcal{A}}}$ — общее число строк в наборе $\widetilde{\mathcal{A}}$. Таким образом, $a(\widetilde{\mathcal{A}}, j)$ пропорционально вероятности неверного классифицирования строки s.

Минимум $a(\widetilde{\mathcal{A}}, j) = 0$ достигается, если множества символов на j-й позиции у разных кластеров не пересекаются. Если, напротив, наборы символов в j-той колонке у всех кластеров совпадают,

$$\forall\, \mathcal{A} \in \widetilde{\mathcal{A}} \quad n_{\mathcal{A}}(\lambda, j) = \begin{cases} 1, & \lambda \in \{\lambda_1, \lambda_2, \ldots, \lambda_l\}, \\ 0, & \text{иначе}, \end{cases}$$

то неоднозначность классифицирования будет максимальной,

$$a(\widetilde{\mathcal{A}}, j) = (n_{\widetilde{\mathcal{A}}} - 1)l = N_{\widetilde{\mathcal{A}}} - l.$$

Если набор $\widetilde{\mathcal{A}}'$ получен из набора $\widetilde{\mathcal{A}}$ объединением подмножеств строк $\mathcal{A}$ и $\mathcal{B} \in \widetilde{\mathcal{A}}$: $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$, $\mathcal{C} \in \widetilde{\mathcal{A}}'$, разность в неоднозначности классифицирования этих двух наборов:

$$a(\widetilde{\mathcal{A}}') - a(\widetilde{\mathcal{A}}) = \sum_{\lambda\in\Lambda}\Bigg(\frac{n_{\mathcal{C}}(\lambda)\big(n_{\widetilde{\mathcal{A}}'}(\lambda) - n_{\mathcal{C}}(\lambda)\big)}{n_{\widetilde{\mathcal{A}}'}(\lambda)} -$$

$$- \frac{n_{\mathcal{A}}(\lambda)\big(n_{\widetilde{\mathcal{A}}}(\lambda) - n_{\mathcal{A}}(\lambda)\big)}{n_{\widetilde{\mathcal{A}}}(\lambda)} - \frac{n_{\mathcal{B}}(\lambda)\big(n_{\widetilde{\mathcal{A}}}(\lambda) - n_{\mathcal{B}}(\lambda)\big)}{n_{\widetilde{\mathcal{A}}}(\lambda)}\Bigg) =$$

$$= \sum_{\lambda\in\Lambda} \frac{1}{n_{\widetilde{\mathcal{A}}}(\lambda)} \big((n_{\widetilde{\mathcal{A}}}(\lambda) - n_{\mathcal{A}}(\lambda) - n_{\mathcal{B}}(\lambda))(n_{\mathcal{A}}(\lambda) + n_{\mathcal{B}}(\lambda)) - n_{\mathcal{A}}(\lambda)(n_{\widetilde{\mathcal{A}}}(\lambda) -$$

$$- n_{\mathcal{A}}(\lambda)) - n_{\mathcal{B}}(\lambda)(n_{\widetilde{\mathcal{A}}}(\lambda) - n_{\mathcal{B}}(\lambda))\big) = -\sum_{\lambda\in\Lambda} \frac{2n_{\mathcal{A}}(\lambda)n_{\mathcal{B}}(\lambda)}{n_{\widetilde{\mathcal{A}}}(\lambda)} \leqslant 0.$$

Таким образом, объединение кластеров уменьшает неоднозначность классифицирования.

5. Свертка векторного критерия качества кластеризации

Итак, для любого разбиения $\widetilde{\mathcal{A}}$ строк матрицы $A \in \Lambda^{M\times N}$ были определены две его характеристики, неоднородность состава и неоднозначность классифицирования:

$$D(\widetilde{\mathcal{A}}) = \sum_{j=1}^{N} d(\widetilde{\mathcal{A}}, j), \quad A(\widetilde{\mathcal{A}}) = \sum_{j=1}^{N} a(\widetilde{\mathcal{A}}, j). \tag{11}$$

Вместе они образуют двухкомпонентный критерий качества кластеризации $Q(\widetilde{\mathcal{A}}) = (D(\widetilde{\mathcal{A}}), A(\widetilde{\mathcal{A}}))$ (см. пример на рис. 1). Чем меньше $D(\widetilde{\mathcal{A}})$, тем более похожи друг на друга строки в пределах одного кластера. Чем меньше $A(\widetilde{\mathcal{A}})$, тем сильнее отличается один кластер от другого. Естественно считать кластеризацию оптимальной, если она обладает обоими этими свойствами. Однако, для произвольной $A \in \Lambda^{N\times M}$ мы, вообще говоря,

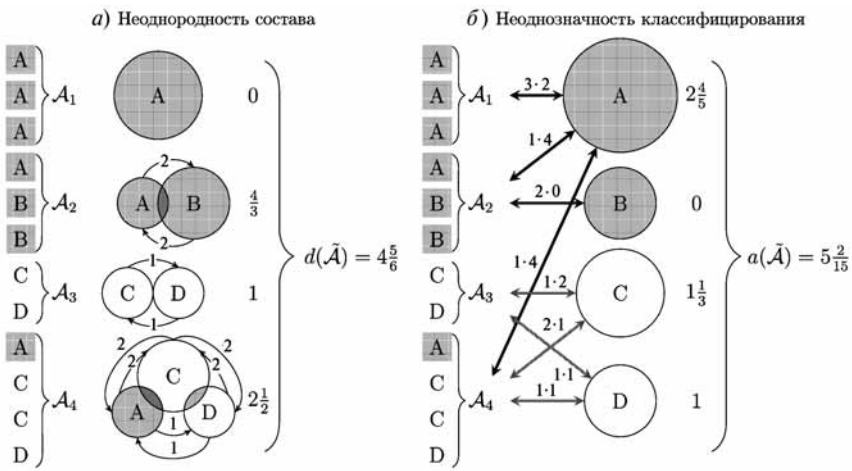

Рис. 1. Графическая интерпретация мер

(*а*) Пример выравнивания

```
FEVGPPAF
NEVGFRGF
KEPTLLGF
NHSFGGSF
KEPTLLGF
FEVGPPAF
FEVGPPAF
SANTCGIF
```

(*б*) $Q(\widetilde{\mathcal{A}})$ для всех возможных кластеризаций

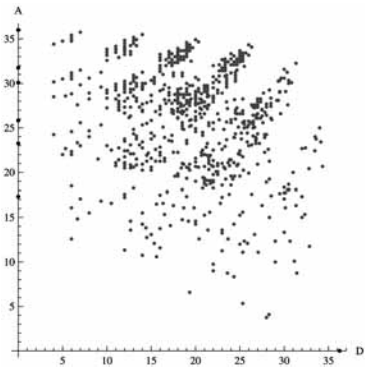

Рис. 2. Пример вычисления $Q(\widetilde{\mathcal{A}})$

можем не найти такого $\widetilde{\mathcal{A}}*$, которое бы минимизировало одновременно $D(\widetilde{\mathcal{A}}*)$ и $A(\widetilde{\mathcal{A}}*)$. Например, подразбиение кластеров на более мелкие части уменьшает $D(\widetilde{\mathcal{A}})$, но увеличивает $A(\widetilde{\mathcal{A}})$. С другой стороны, объединение кластеров уменьшает $A(\widetilde{\mathcal{A}})$ и увеличивает $D(\widetilde{\mathcal{A}})$ (см. рис. 2).

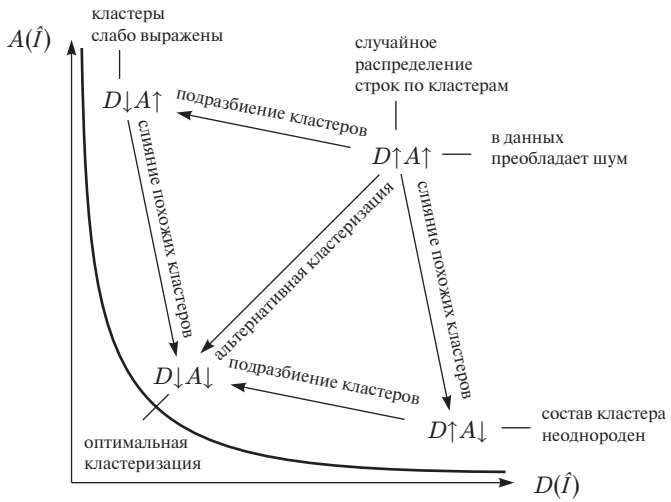

Рис. 3. Множество Парето для $Q(\widetilde{\mathcal{A}})$ и диаграмма преобразования кластеризаций

Все кластеризации, такие что ни $D(\widetilde{\mathcal{A}})$, ни $A(\widetilde{\mathcal{A}})$ не могут быть уменьшены без увеличения другой меры, являются *Парето оптимальными* кластеризациями $(\mathrm{PO}(\mathfrak{A}))$ (см. рис. 3).

Для выбора *единственной* оптимальной кластеризации необходимы дополнительные соображения. Одним из возможных решений является рассмотрение некоторой *свертки* функций $D(\widetilde{\mathcal{A}})$ и $A(\widetilde{\mathcal{A}})$. Та оптимальная по Парето кластеризация $\widetilde{\mathcal{A}}* \in \mathrm{PO}(\mathfrak{A})$, которая будет доставлять минимум этому функционалу, будет считаться оптимальной. Функция

$$q_w(\widetilde{\mathcal{A}}) = E^2(\widetilde{\mathcal{A}}) + wC^2(\widetilde{\mathcal{A}}) \quad w > 0 \tag{12}$$

является одним из примеров такой свертки. По определению, $D(\widetilde{\mathcal{A}}) \leqslant M$ и $A(\widetilde{\mathcal{A}}) \leqslant M$, поэтому можно положить вес w равным 1. Точка минимума

$$\widetilde{\mathcal{A}}* = \mathrm{argmin}_{\widetilde{\mathcal{A}} \in \mathrm{PO}(\mathfrak{A})}\, q_1(\widetilde{\mathcal{A}}) = \mathrm{argmin}_{\widetilde{\mathcal{A}} \in \mathfrak{A}}\, q_1(\widetilde{\mathcal{A}})$$

будет являться проекцией начала координат пространства $D - A$ на образ оптимальных по Парето кластеризаций.

Литература

1. *Журавлев Ю.*, *Рязанов В.*, *Сеньков О.* «Распознавание». Математические методы. Программная система. Практические применения. М.: Фазис, 2006.
2. Cluster analysis and display of genome-wide expression patterns / M. B. Eisen, P. T. Spellman, P. O. Browndagger, D. Botstein // Proceedings of the National Academy of Sciences. 1998. V. 95. P. 14863–14868.
3. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs / S. F. Altschul, T. L. Madden, A. A. Schäffer et al. // Nucleic Acids Research. 1997. V. 25. № 17. P. 3389–3402.
4. *Henikoff S.*, *Henikoff J. G.* Amino acid substitution matrices from protein blocks // Proceedings of the National Academy of Sciences of the United States of America. 1992. V. 89. P. 10915–10919.
5. *Reva B.*, *Antipin Y.*, *Sander C.* Determinants of protein function revealed by combinatorial entropy optimization. // Genome Biology. 2007. V. 8. № 11. P. R232 (http://genomebiology.com/2007/8/11/R232).
6. *Thompson J. D.*, *Higgins D. G.*, *Gibson T. J.* CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice // Nucleic Acids Research. 1994. V. 22. № 22. P. 4673–4680.