Color-coding Similarity in Sequence Alignment

Asimakis Kydros
3881
asimakis@csd.auth.gr

December 10, 2023

Purpose of Color Schemes

In sequence alignment, it's useful to color the amino acids in a way that predicts their similarity, so we can check with a glance the estimated alignment.

Typically, this is done manually, by experienced scientists, using colors according to pre-determined physical properties, such as charge.

This isn't perfect multiple-fold, as two amino acids having a similar property doesn't always ensure that their alignment will be optimal in the given context. Also, we would like an automated solution for every situation.

Substitution Matrix Schema

A better way to do this is to use the given substitution matrix and a chosen color space, and define through them a scoring function. Then, optimize that to approximate the optimal.

Given the substitution matrix M, define

$$D' \ni \left(D'_{ij} = \begin{cases} \frac{M_{ii}-M_{ij}+M_{jj}-M_{ji}}{2}, & j \le i \\ 0, & j > i \end{cases} \right) \xrightarrow{\text{scaled to average 1}} D$$

and the chosen color space *CIE L*a*b**, define

$$C \ni \left(C_{ij} \approx \sqrt{(L_i^* - L_j^*)^2 + (a_i^* - a_j^*)^2 + (b_i^* - b_j^*)^2} \right)$$

These are triangular matrices that define the pairwise distance and color variance respectively.

Substitution Matrix Schema

We can now define the score function

$$S_T = S_H + S_C$$

where

$$S_H = \sum_{ij}(f_s C_{ij} - D_{ij})^2$$

$$S_C = \frac{f_C}{\langle C \rangle}$$

$$f_s = \frac{\langle D \rangle}{\langle C \rangle}$$

$\langle \cdot \rangle$ the arithmetic mean

f_C a given hyperparameter

which essentially defines the total square error of the scaled color variance to distance for each pair, S_H, biased by the contrast enabler S_C.

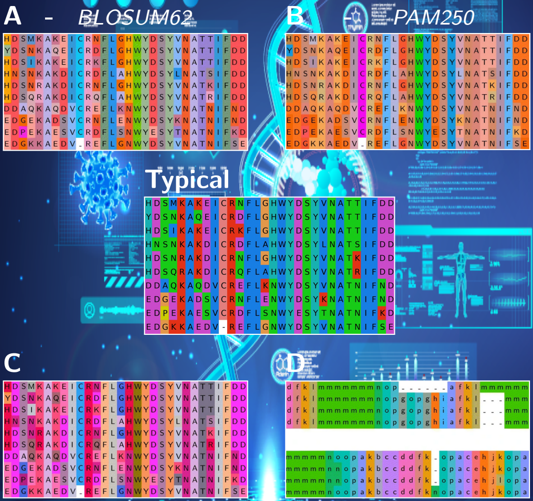
This function describes a (generally) non-convex problem that adequately combines the data into the desired effect: coloring based on distance. It can be optimized with Simulated Annealing.

Results

The figure grid compares this algorithm’s outputs (A, B, C and D) to some given color schema made manually.

We can see that the approaches agree on color similarity, but differ on color dissimilarity; the palette on A and B is more gradient and nuanced than the one on ’Typical’, which has very sharp changes.

The versatility of the method is also emphasized; Figure C is adapted to red-green colorblindness by removing green as an option. Figure D uses a different alphabet altogether, the protein blocks.



A - BLOSUM62

```
H D S M K A K E I C R N F L G H W Y D S Y V N A T T I F D D
Y D S N K A Q E I C R D F L G H W Y D S Y V N A T T I F D D
H D S I K A K E I C R K F L G H W Y D S Y V N A T T I F D D
H N S N K A K D I C R D F L A H W Y D S Y L N A T S I F D D
H D S N R A K D I C R N F L G H W Y D S Y V N A T K I F D D
H D S Q R A K D I C R Q F L A H W Y D S Y V N A T R I F D D
D D A Q K A Q D V C R E F L K N W Y D S Y V N A T N I F N D
E D G E K A D S V C R N F L E N W Y D S Y K N A T N I F N D
E D P E K A E S V C R D F L S N W Y E S Y T N A T N I F K D
E D G K K A E D V - R E F L G N W Y D S Y V N A T N I F S E
```

B PAM250

```
H D S M K A K E I C R N F L G H W Y D S Y V N A T T I F D D
Y D S N K A Q E I C R D F L G H W Y D S Y V N A T T I F D D
H D S I K A K E I C R K F L G H W Y D S Y V N A T T I F D D
H N S N K A K D I C R D F L A H W Y D S Y L N A T S I F D D
H D S N R A K D I C R N F L G H W Y D S Y V N A T K I F D D
H D S Q R A K D I C R Q F L A H W Y D S Y V N A T R I F D D
D D A Q K A Q D V C R E F L K N W Y D S Y V N A T N I F N D
E D G E K A D S V C R N F L E N W Y D S Y K N A T N I F N D
E D P E K A E S V C R D F L S N W Y E S Y T N A T N I F K D
E D G K K A E D V - R E F L G N W Y D S Y V N A T N I F S E
```

Typical

```
H D S M K A K E I C R N F L G H W Y D S Y V N A T T I F D D
Y D S N K A Q E I C R D F L G H W Y D S Y V N A T T I F D D
H D S I K A K E I C R K F L G H W Y D S Y V N A T T I F D D
H N S N K A K D I C R D F L A H W Y D S Y L N A T S I F D D
H D S N R A K D I C R N F L G H W Y D S Y V N A T K I F D D
H D S Q R A K D I C R Q F L A H W Y D S Y V N A T R I F D D
D D A Q K A Q D V C R E F L K N W Y D S Y V N A T N I F N D
E D G E K A D S V C R N F L E N W Y D S Y K N A T N I F N D
E D P E K A E S V C R D F L S N W Y E S Y T N A T N I F K D
E D G K K A E D V - R E F L G N W Y D S Y V N A T N I F S E
```

C

```
H D S M K A K E I C R N F L G H W Y D S Y V N A T T I F D D
Y D S N K A Q E I C R D F L G H W Y D S Y V N A T T I F D D
H D S I K A K E I C R K F L G H W Y D S Y V N A T T I F D D
H N S N K A K D I C R D F L A H W Y D S Y L N A T S I F D D
H D S N R A K D I C R N F L G H W Y D S Y V N A T K I F D D
H D S Q R A K D I C R Q F L A H W Y D S Y V N A T R I F D D
D D A Q K A Q D V C R E F L K N W Y D S Y V N A T N I F N D
E D G E K A D S V C R N F L E N W Y D S Y K N A T N I F N D
E D P E K A E S V C R D F L S N W Y E S Y T N A T N I F K D
E D G K K A E D V - R E F L G N W Y D S Y V N A T N I F S E
```

D

```
d f k l m m m m m m m n o p - - - - - - a f k l m m m m m m
d f k l m m m m m m m n o p g o p g h i a f k l - - - m m m
d f k l m m m m m m m n o p g o p g h i a f k l - - - m m m
d f k l m m m m m m m n o p g o p g h i a f k l - - - m m m

m m m m m n o o p a k b c c d d f k _ o p a c e h j k o p a
m m m m m n n o p a k b c c d d f k _ o p a c e h j k o p a
m m m m m n n o p a k b c c d d f k _ o p a c e h j l o p a
m m m m m n o o p a k b c c d d f k n o p a c e h j k o p a
```

Conclusions

This tool does a terrific job projecting the relationships between amino acids anchored on molecular evolution rather than anything else, thus tying them to the problem of sequence alignment.

It enables the automatic and independent calculation of sequence alignment heat maps, without the need for research of physical characteristics and unnecessary modeling.

Its generic definition also allows it to be applied to fields other than bioscience, be it for understanding exotic alphabets, cryptographic concepts and much more.

References

Principal paper:

- Kunzmann, P.; Mayer, B.E.; Hamacher, K. (2020) *”Substitution matrix based color schemes for sequence alignment visualization”*

Secondary sources:

- The course’s slides and notes.
- Wikipedia *”Multiple sequence alignment”*