Μελέτη αναπαραστάσεων γονιδιωματικών ακολουθιών σε προβλήματα ταξινόμησης

ΠΜΣ Βιοπληροφορικής

Μαρίνα-Αγάπη Αθανασούλη

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
ΕΚΕΦΕ Δημόκριτος

Σκοπός

- Σύντομη παρουσίαση διαφορετικών αναπαραστάσεων γονιδιωματικών ακολουθιών
- Αξιολόγηση των συγκεκριμένων αναπαραστάσεων όσον αφορά τη χρήση τους για την ταξινόμηση ακολουθιών DNA
- Εύρεση και ταξινόμηση CNE's (Conserved Non-Expressed regions) εντός και μεταξύ συγκεκριμένων ειδών σε συνέχεια του Polychronopoulos *et al.*, 2014

Ερευνητικό ενδιαφέρον

- Ανάλυση DNA ακολουθιών με επεξεργασία ψηφιακού σήματος (Digital Signal Processing)
- Απαραίτητη η εύρεση και χρήση τρόπων αναπαράστασης των ακολουθιών υψηλής ακρίβειας
- Πληθώρα αναπαραστάσεων→επιλογή κατάλληλης, ανάλογα με τον πειραματικό σκοπό
- Βασική απαίτηση: η απεικόνιση των βιολογικών ιδιοτήτων των ακολουθιών και η αποφυγή μαθηματικών συσχετίσεων που δεν υφίστανται

Είδη αναπαραστάσεων

1. **Αριθμητικές:** μετατροπή της γονιδιωματικής (συμβολικής) ακολουθίας σε αριθμητική

   Δύο υποκατηγορίες:

   - Σταθερές τιμές: μετατροπή σε σειρά αυθαίρετων αριθμητικών ακολουθιών
   - Βάσει φυσικοχημικών ιδιοτήτων

2. Γραφικές

3. Στατιστικά μοντέλα

4. Μαρκοβιανά μοντέλα

Αριθμητικές αναπαραστάσεις

Fixed-mapping values(1)

Voss, 1992: Αναπαράσταση Voss

$X_n = 1$ for $S(n) = X$ $X_n = 0$ for $S(n) \neq X$ X_n applies to any of C_n, G_n, A_n, T_n	$C_n = [1, 0, 0, 0]$ $G_n = [0, 1, 0, 0]$ $A_n = [0, 0, 1, 0]$ $T_n = [0, 0, 0, 1]$

, όπου $S_n = CGAT$

Silverman & Linker, 1986 και Cristea, 2002: Τετράεδρο

$x_r(n) = \frac{\sqrt{2}}{3}[2T_n - C_n - G_n]$	$x_r(n) = \frac{\sqrt{2}}{3}[-1,-1,0,2]$
$x_g(n) = \frac{\sqrt{6}}{3}[C_n - G_n]$	$x_g(n) = \frac{\sqrt{6}}{3}[1,-1,0,0]$
$x_b(n) = \frac{1}{3}[3A_n - T_n - C_n - G_n]$	$x_b(n) = \frac{1}{3}[-1,-1,3,-1]$

Fixed-mapping values(2)

Cristea, 2002: Complex

A = 1+j, C = -1+j, G = -1-j, T = 1-j	[-1+j, -1-j, 1+j, 1-j]

, $S_n = CGAT$

Cristea, 2002: Integer

A = 2, C = 1, G = 3, T = 0	[1, 3, 2, 0]

Chakravarthy *et al.*, 2004: Real Number

A = -1.5, C = 0.5, G = -0.5, T = 1.5	[0.5, -0.5, -1.5, 1.5]

Physico-Chemical properties

Achuthsankar & Sreenadhan, 2006: Electron-Ion Interaction potential (EIIP)

A= 0.1260, C= 0.1340, G= 0.0806, T= 0.1335	[0.1340, 0.0806, 0.1260, 0.1335]

, $S_n = CGAT$

Holden *et al.*, 2007:Atomic Number

A = 70, C = 58, G = 78, T = 66	[58, 78, 70, 66]

Akhtar *et al.*, 2007: Paired Numeric

A or T = 1, C or G = -1	P_{1n} = [-1, -1, 1, 1]
	P_{2n} = [-1, -1, 0, 0] & [0, 0, 1, 1]

Γραφικές Αναπαραστάσεις

Graphical(1)

Hamori & Rushkin, 1983: H-curve, πρώτη 3D αναπαράσταση για DNA ακολουθίες

Zhang & Zhang, 1994: Z-curve

$$x_n = (A_n + G_n) - (C_n + T_n) \equiv R_n - Y_n \qquad x_n = [-1, 0, 1, 0]$$
$$y_n = (A_n + C_n) - (G_n + T_n) \equiv M_n - K_n \qquad y_n = [1, 0, 1, 0]$$
$$z_n = (A_n + T_n) - (C_n + G_n) \equiv W_n - S_n \qquad z_n = [-1, -2, -1, 0]$$

, $S_n = CGAT$

Nandy *et al.*, 1994:

- 2D αναπαράσταση βασισμένη στην H-curve
- Ανάθεση των πουρινών στον άξονα x και των πυριμιδινών στον άξονα y

Randic *et al.*

- 3D (2000) και 2D (2003) αναπαράσταση
- Εναλλακτική προσέγγιση: αριθμητικός χαρακτηρισμός του DNA

Graphical(2)

3-D αναπαραστάσεις

1. **Qi & Fan, 2007:** PN-curve
   Χαρτογράφηση βάσει ζευγών νουκλεοτιδίων π.χ. εάν $g_i g_{i+1} = AA(g_i g_{i+1}) = (1, (AA)_i, i)$ κ.ο.κ
2. **Yu *et al.*, 2009:** TN-curve
   Αναπαράσταση τρινουκλεοτιδίων με σετ συντεταγμένων (x,y) και χαρτογράφηση ακολουθίας S: $(S) = (s_1 s_2 s_3)...(s_i s_{i+1} s_{i+2})$, όπου $(s_i s_{i+1} s_{i+2}) = \{(x_i, y_i, i)\}$

4-D αναπαραστάσεις

1. **Tang *et al.*, 2009:** Βασισμένη στη Z-curve
2. **Tan *et al.*, 2014:** Βασισμένη στην 2-D αναπαράσταση που πρότεινε ο Randic

Graphical(3)

Jeffrey, 1990: Αναπαράσταση παιγνίου χάους (Chaos Game Representation)

- **Deschavanne *et al.*, 1999:** Χρήση ως γονιδιωματικές υπογραφές
- **Karamichalis, 2016**: Προσθετικές γονιδιωματικές υπογραφές

Qi *et al.*, 2011: Κωδικοποίηση Huffman

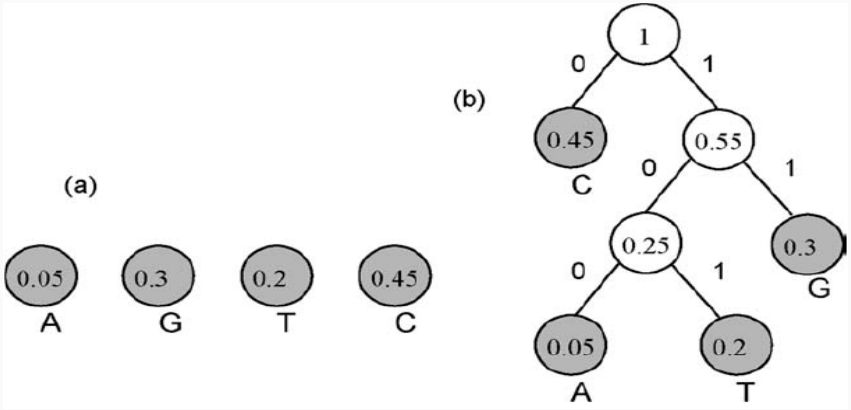

Figure 1: (a) Κόμβοι-φύλλα του δυαδικού δένδρου. (b) Δημιουργία του δένδρου

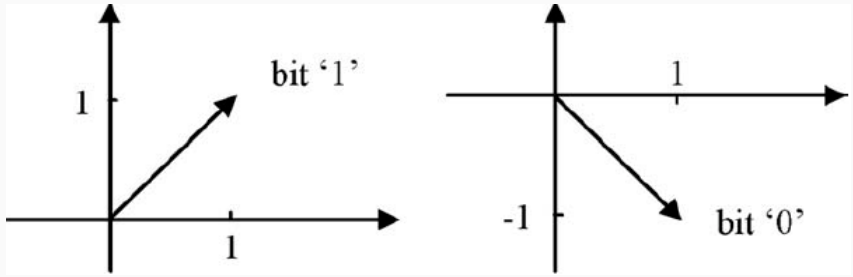

Figure 2: Μοναδιαία διανύσματα για τα bits

Στατιστικά μοντέλα

Στατιστικά μοντέλα

Aires-de-Sousa & Aires-de-Sousa, 2003: Virtual Potentials

- Εικονικό δυναμικό μιας βάσης: $VP_{XY} = \frac{1}{d_{XY}}$
- Αναπαράσταση DNA ακολουθιών με το ιστόγραμμα των εικονικών δυναμικών (SEQPREP)

O'Neill & Erill, 2016: Κατανομές Maximum Entropy και Truncated Uniform

- Κατασκευή κατανομών βάσει του Information Content συγκεκριμένων μοτίβων

Μαρκοβιανά μοντέλα

Markov Models

Liu *et al.*, 2007: Χρήση Μαρκοβιανών μοντέλων ανώτερης τάξης και μεταβλητής τάξης

- Μαρκοβιανό μοντέλο ανώτερης τάξης: με minimum description length και Akaike's Information Criterion
- Μαρκοβιανό μοντέλο μεταβλητής τάξης: με δέντρο πιθανοτήτων (Probability Suffix Tree)

Pinho *et al.*, 2011: Χρήση πολλαπλών "ανταγωνιστικών" πεπερασμένων Μαρκοβιανών μοντέλων

- Ανταγωνιστικά μοντέλα διαφορετικών τάξεων→επικράτηση του καλύτερου

Λόγοι επιλογής

- Αντιπροσωπευτικό δείγμα από διαφορετικούς τύπους αναπαραστάσεων
- Σύγκριση των διαφορετικών προσεγγίσεων που χρησιμοποιούν
- Αξιολόγηση της απόδοσης τους στους ερευνητικούς σκοπούς που χρησιμοποιούνται συνήθως

Βιβλιογραφία


Akhtar, M., Epps, J., and Ambikairajah, E. (2007).
On DNA Numerical Representations for Period-3 Based Exon Prediction.
In *2007 IEEE International Workshop on Genomic Signal Processing and Statistics*, pages 1–4.


Kwan, H. K. and Arniker, S. B. (2009).
Numerical representation of DNA sequences.
In *2009 IEEE International Conference on Electro/Information Technology*, pages 307–310.


Mendizabal-Ruiz, G., Roman-Godinez, I., Torres-Ramos, S., Salido-Ruiz, R. A., and Morales, J. A. (2017).
On DNA numerical representations for genomic similarity computation.

[Kwan and Arniker, 2009] [Akhtar et al., 2007]
[Mendizabal-Ruiz et al., 2017]

Questions?