iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition && My own idea for prediction hotspot && Repeating Lin kui's work

Dong Chuan

1. Background
2. Materials and Methods
3. Results and Discussion
4. Summary

1.Background

- Genetic recombination is an important biological process
- In the past decades, several global mapping studies have been performed to map doublestrand breaks sites on chromosomes in yeast to determine the distribution pattern of recombination regions across genome;
- it is highly desired to develop reliable automated methods for timely identifying the recombination spots(hot spots) and non-recombination spots(cold spots);

- The existing computational algorithm for recombination spots prediction was based on the nucleotide sequence contents, and sequence-order effect is not taken into account;
- Some predict methods condidering sequence order such as PseAAC approach introducing into computational proteomics and encouraged by the successes of this method, the author propose a novel feature vector, called pseudo dinucleotide composition’(PseDNC) to predict the recombination spots.

2.Materials and Methods

- 2.1 Benchmark data set

  a. The positive and negative set came from Liu et al(see Ref 6 in this paper), which 490 recombination hotspots and 591 recombination coldspots(train data set);

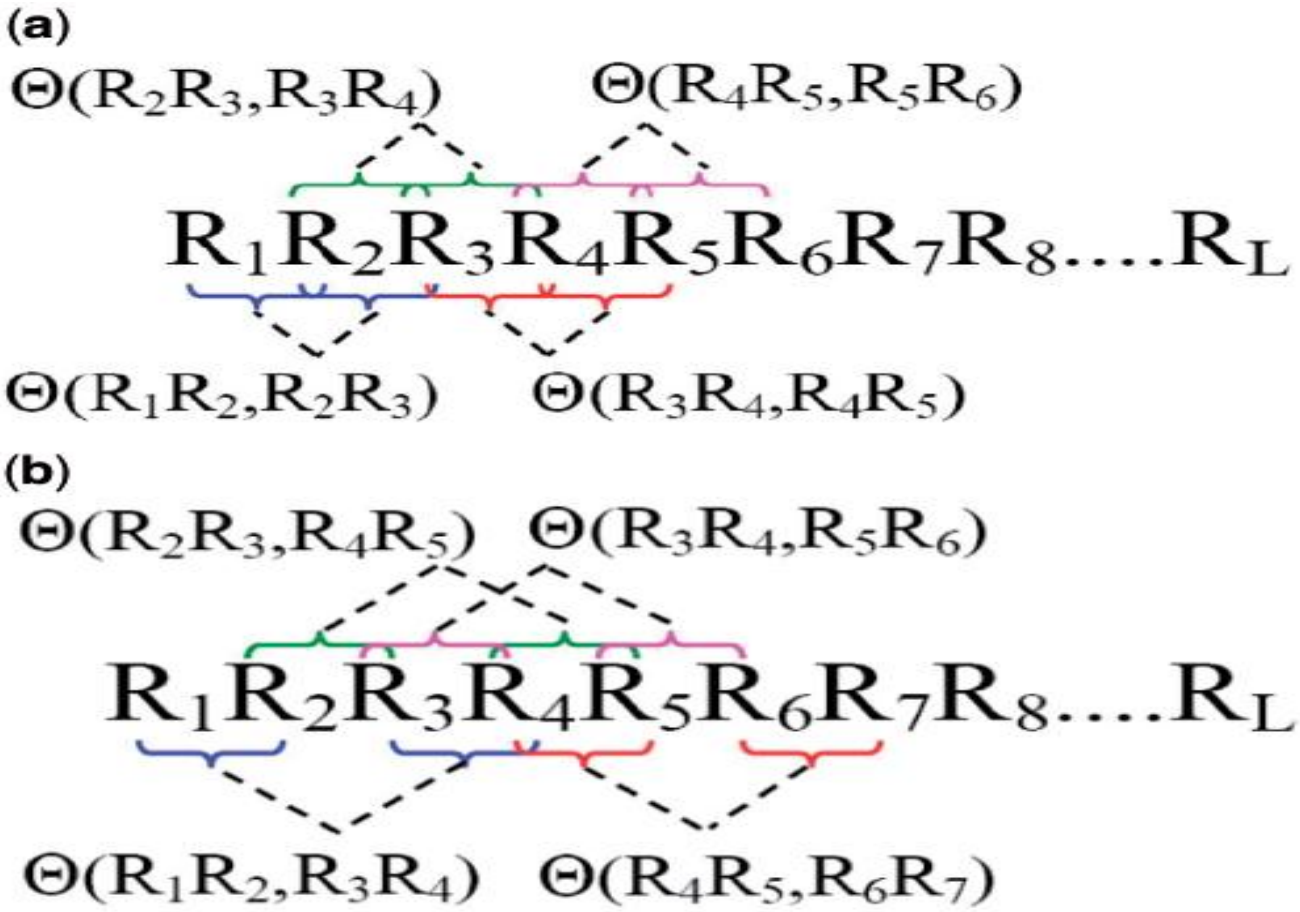
  b. Come from Pan et al's data (See ref 58 in this paper)(validate the reliability of the method).

- 2.2 PseDNC
- 2 mean value+an important formula

$$\Theta(\mathrm{R}_i\mathrm{R}_{i+1},\mathrm{R}_j\mathrm{R}_{j+1}) = \frac{1}{\mu}\sum_{u=1}^{\mu}\left[P_u(\mathrm{R}_i\mathrm{R}_{i+1}) - P_u(\mathrm{R}_j\mathrm{R}_{j+1})\right]^2$$

$$
\begin{cases}
\theta_1 = \frac{1}{L-2}\sum_{i=1}^{L-2}\Theta(R_iR_{i+1},R_{i+1}R_{i+2}) \\
\theta_2 = \frac{1}{L-3}\sum_{i=1}^{L-3}\Theta(R_iR_{i+1},R_{i+2}R_{i+3}) \\
\theta_3 = \frac{1}{L-4}\sum_{i=1}^{L-4}\Theta(R_iR_{i+1},R_{i+3}R_{i+4}) \qquad (\lambda < L) \\
\cdots\cdots \\
\theta_\lambda = \frac{1}{L-1-\lambda}\sum_{i=1}^{L-1-\lambda}\Theta(R_iR_{i+1},R_{i+\lambda}R_{i+\lambda+1})
\end{cases}
$$

where $\Theta_1$is called the first-tier correlation factor that reflects the sequence-order correlation between all the most contiguous dinucleotide along a DNA sequence; Θ_2, the second-tier correlation factor between all the second most contiguous dinucleotide and so on (see next figure)



Further explanation for Θ : Step of windows. R_iR_{i+1} the standard conversion DNA local physical structural properties, such as twist, tilt, roll, shift, slide and rise

- Novel vector called D tranformed by Θ_λ and f：
  $D=[d_1\ d_2\ \ldots\ldots\ d_{16}\ d_{16+\lambda}]$,
  where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (17 \le k \le 16+\lambda) \end{cases}$$

The novel vector D was called PseDNC

- 2.3 Other methods
- SVM

  Radial basis kernel function;

  Grid search approach:C=32, γ=0.5;

  Using D as the input vectors:λ=3 and ω=0.05.

3.Results and Discussion

Table 3. A comparison of between iRSpot-PseDNC with the existing method

Predictor	Test method	Sn (%)	Sp (%)	Acc (%)	MCC
iRSpot-PseDNC[a]	Jackknife	73.06	89.49	82.04	0.638
	5-fold cross	81.63	88.14	85.19	0.692
IDQD[b]	5-fold cross	79.40	81.00	80.30	0.603

[a]The parameters used: $\lambda = 3$ and $w = 0.05$ for Equation 9; $C = 32$ and $\gamma = 0.5$ for the LIBSVM operation engine (47).
[b]From Liu *et al.* (6).

used iRSpot-PseDNC to identify the 452 experimentally annotated recombination hotspots by Pan et al. Results obtained by iRSpot-PseDNC the overall success rate was 76.77%.

4.Summary

- The advantage of pseudo dinucleotide: Firstly, The sequence-order effect was taken into account; Secondly, some physicochemical parameters were also contained; Thirdly, This methods can reflect the biosequence from microcosmic. Due to the physicochemical parameters too much, how can we seclect the best parameters for this methods. In addition, how to confirm the best ω, ɵ is also a question

My own idea for prediction hotspot

- An idea called physiochemical mass center copying from the field of physics was introduced to analysis the biosequence:

ASTGHWRKLMNE

i

$$x_i = i - 1$$

$$y_i = L - i$$

Analogy with particle system in physics

ASTKCMEWRKLMNDFHIQ
VY

…… …. … . . . . .

…… …… ….

(xi,yi)

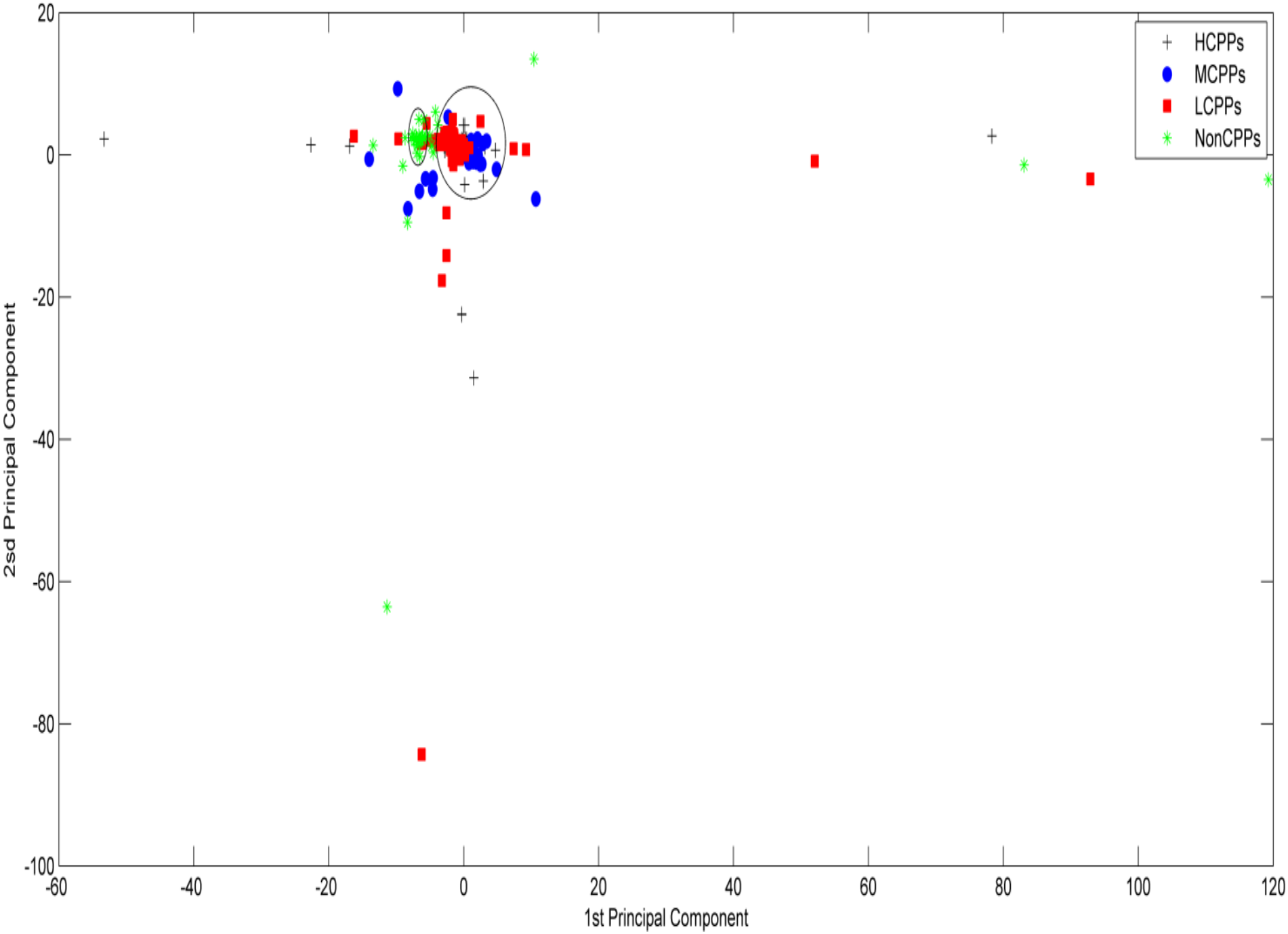
(xi,yi)

Some physicochemical parameter

m1, m2, m3, m4

$$x_c = \frac{\sum_{i=1}^{I} PC_i x_i}{\sum_{i=1}^{I} PC_i}$$

$$y_c = \frac{\sum_{i=1}^{I} PC_i y_i}{\sum_{i=1}^{I} PC_i}$$


HCPPs
MCPPs
LCPPs
NonCPPs
2sd Principal Component
1st Principal Component
20
0
-20
-40
-60
-80
-100
-60
-40
-20
0
20
40
60
80
100
120

- physiochemical mass center can reflect biosequence from their whole. If combine the two methods to predict hotspot can their predict accuracy can be improved?????

Repeating Lin kui's work