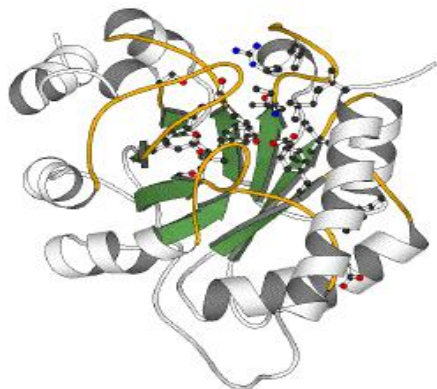
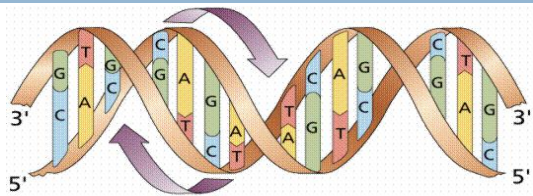


# Z曲线在原核生物基因预测中的应用

# 中心法则



DNA

transcription

RNA

translation

Protein

CCTGAGCCAACTATTGATGAA



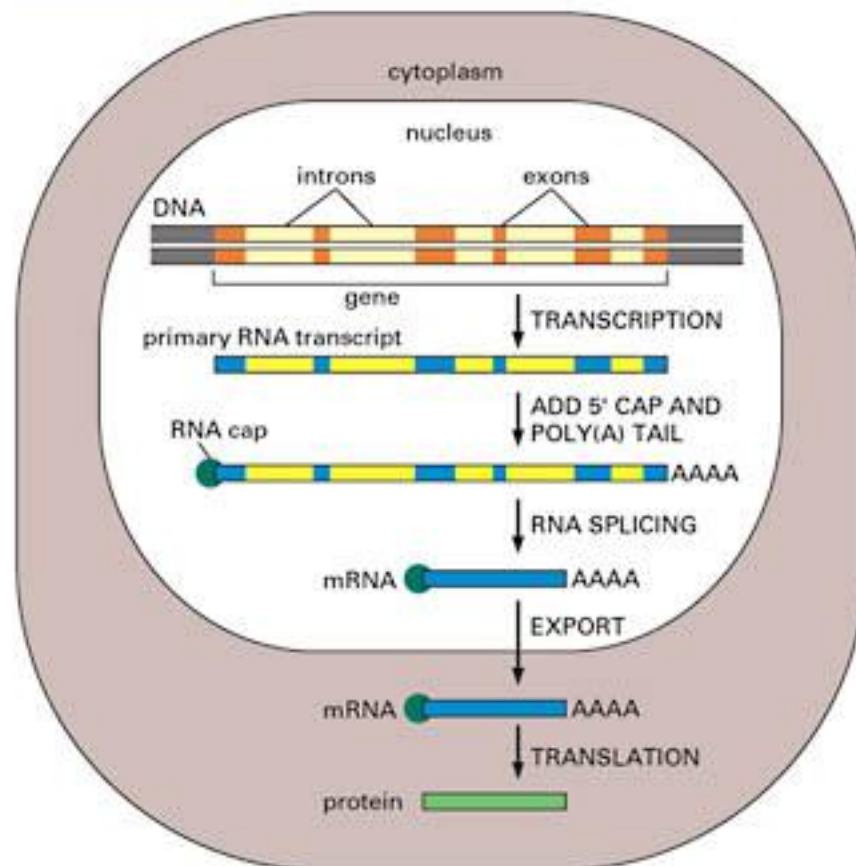
CCUGAGCCAAACUAUUGAUGAA



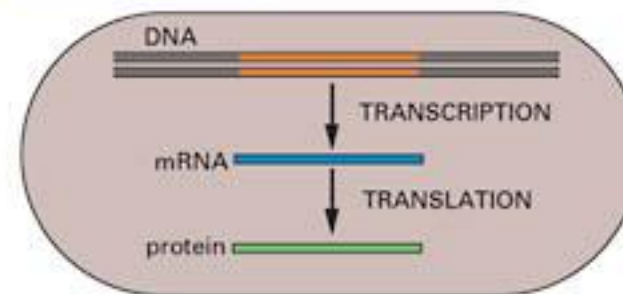
PEPTIDE

# 中心法则

(A) EUCARYOTES



(B) PROCARYOTES



# 基因的类型

- 编码蛋白的基因

- ▣ 原核生物

- 没有内含子
    - 结构简单

- ▣ 真核生物

- 内含子
    - 结构复杂

# 原核与真核生物基因预测的比较

## 原核生物基因

- 基因组较小  $0.5 - 10 \cdot 10^6$  bp
- 编码区密度大 ( $>90\%$ )

- 没有内含子

↓  
--基因预测相对容易，准确率在99%左右

### 存在的问题：

- ORFs的重叠
- 短基因的预测
- 转录开始位点的预测

## 真核生物基因

- 基因组较大  $10^7 - 10^{10}$  bp
- 编码区密度小 ( $<50\%$ )

- 内含子/外显子结构

↓  
--基因预测的水平较低，准确率在50%左右

### 存在的问题：

- 很多

# 基因识别的常用方法

- 同源性方法
- 基因组的比较
- 从头预测的方法
- 综合方法

# 同源性的方法

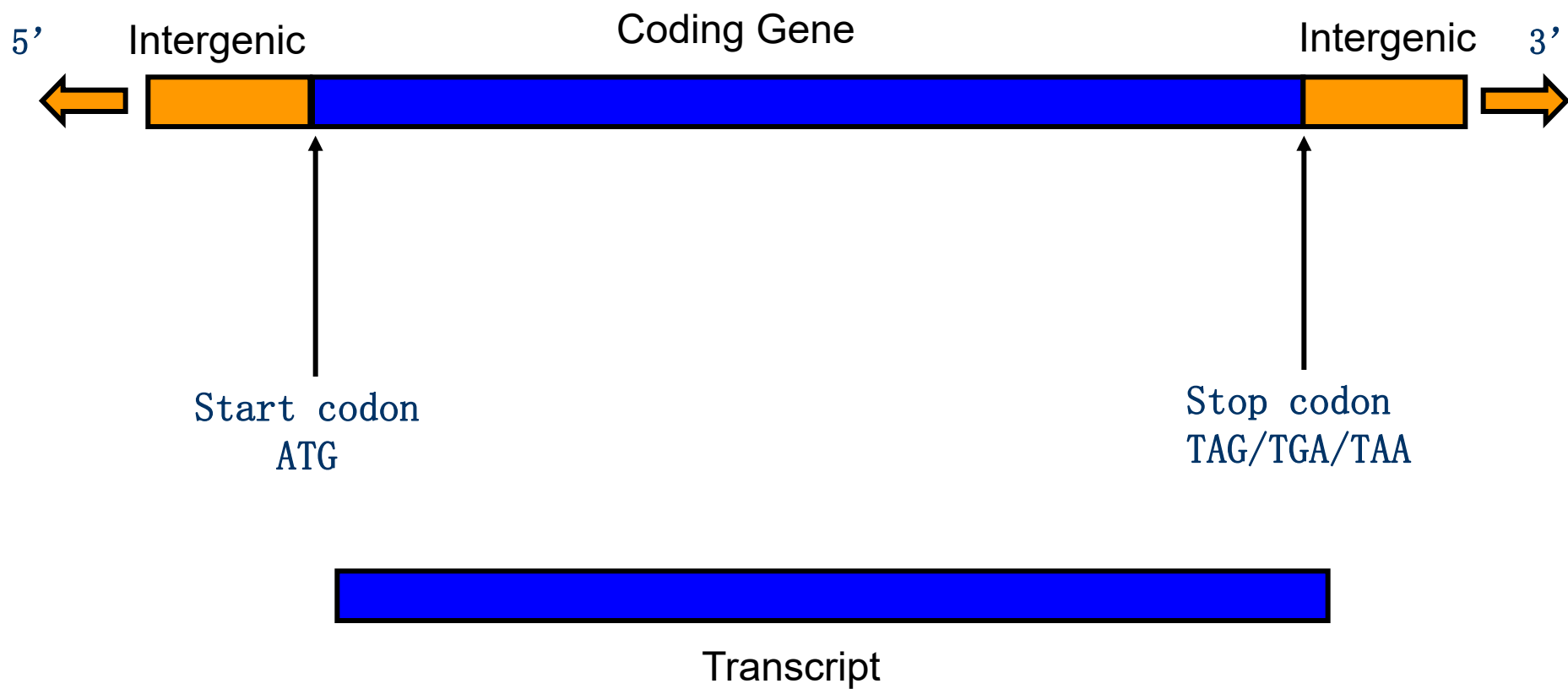
- 具有功能的序列相对保守
- 使用局部比对的方法 (Smith-Waterman algo, BLAST, FASTA) 搜索protein, cDNA, 和 EST 数据库
- 不能识别不在数据库中的基因 (能够发现 ~50% 新基因)
- 序列相似的阈值难以确定

# 基因组比较的方法

- 编码区一般会比非编码区保守
- 得到同源区域的比对
- 认为在基因组间保守的区域应为编码区
- 序列相似的阈值难以确定



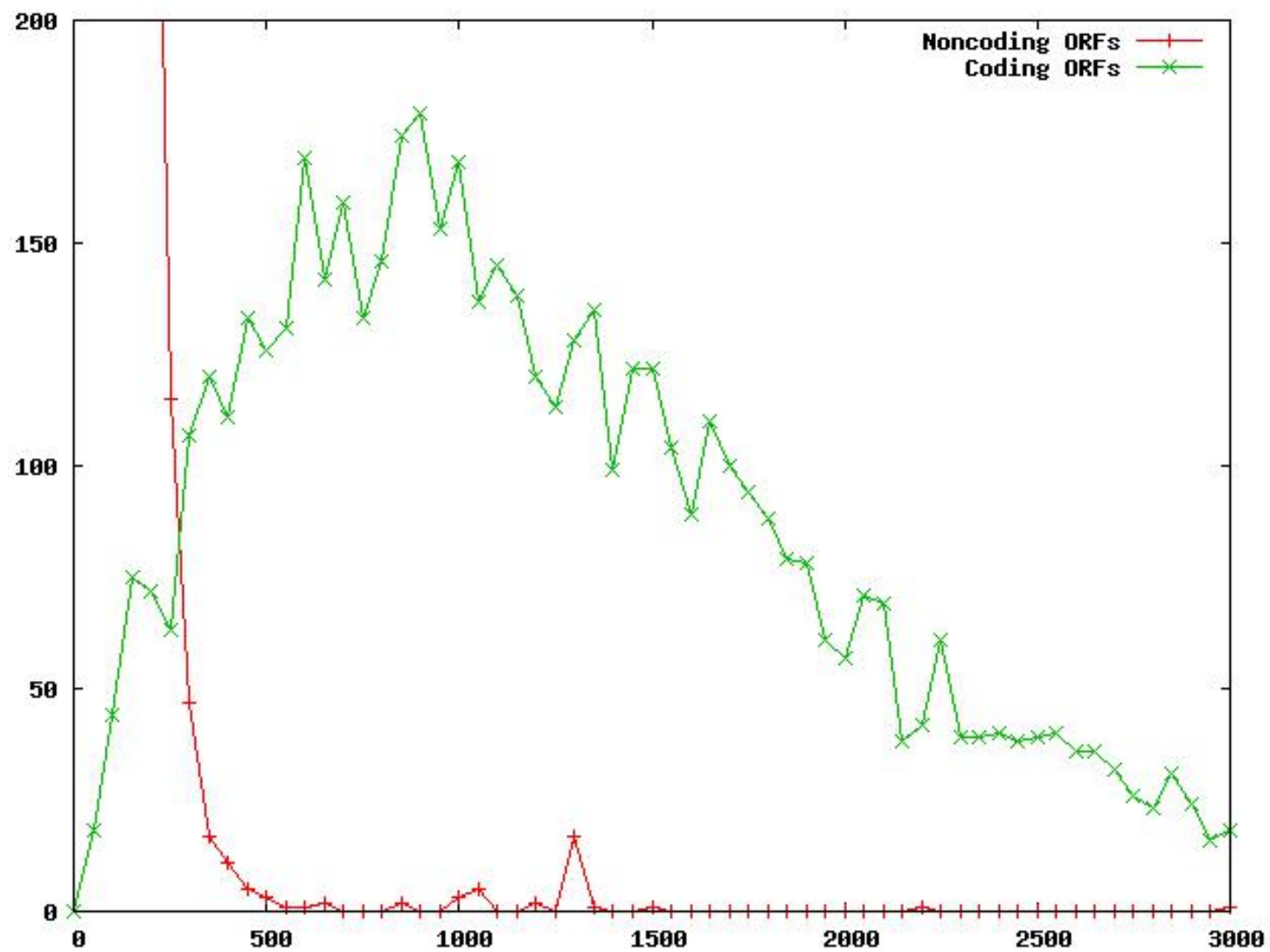
# 原核生物基因结构



# ORF搜索

- Open Reading Frame(ORF): 以起始密码子（ATG,CTG,GTG,TTG）开始，以同相位的终止密码子结束，并且中间没有终止密码子的序列片段
- ORF搜索
  - ▣ 在序列上搜索(ORFs)
  - ▣ 需要在6种不同的“相位”上搜索
  - ▣ 通过ORFs的特征来判断是否为基因，比如，长的ORFs一般为基因。
  - ▣ 一个DNA序列有3个正向阅读框，3个反向阅读框，在6个阅读框中，通常只有一个正确翻译产生蛋白质，其他5个为非编码的。

# 酵母(Yeast)ORF的分布



# Z-曲线

- 由天津大学张春霆院士提出的**DNA**表示的新方法
- **Z**曲线和**DNA**序列是一一对应的，可以由**Z**曲线得到**DNA**序列，同样也可以由**DNA**序列得到**Z**曲线

# Z变换与Z曲线

- 对于一个长度为N的DNA序列，Z曲线定义为：

$$\begin{cases} x_n = (A_n + G_n) - (C_n + T_n), \\ y_n = (A_n + C_n) - (G_n + T_n), \quad n = 0, 1, 2, \dots, N \\ z_n = (A_n + T_n) - (C_n + G_n), \end{cases}$$

- 其中,  $A_n, G_n, C_n, T_n$  分别表示从第一个碱基到第n个碱基的子序列中四种碱基A,C,G,T的各自出现的次数. 定义

$$A_0 = 0, G_0 = 0, C_0 = 0, T_0 = 0$$

从而

$$x_0 = 0, y_0 = 0, z_0 = 0$$

Z曲线为从原点出发的折线

# Z变换的逆变换

- 给定一条Z曲线的坐标, 它所表示的DNA序列可以用Z变换的逆变换重构出来:

$$\begin{pmatrix} A_n \\ C_n \\ G_n \\ T_n \end{pmatrix} = \frac{n}{4} \times \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \frac{1}{4} \times \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \times \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix}, \quad n = 0, 1, 2, \dots, N$$

- 其中  $A_n + G_n + C_n + T_n = n$

# Z曲线的意义

- Z曲线在3个坐标轴上的投影形成了它的三个独立的分量。Z曲线的三个分量分别描述了DNA序列的三种独立的碱基分布：
- Xn分量表示嘌呤/嘧啶(A+G/C+T)沿序列的分布
- Yn分量表示氨基/酮基(A+C/G+T)沿序列的分布
- Zn分量表示强/弱氢键(A+T/C+G)沿序列的分布

# Z曲线与原核基因组

- 使用Z曲线分析原核基因组中ORFs的碱基分布
- 数据集Set1：
  - ▣ *S. coelicolor* A3(2)基因组 (GenBank AL645882) , 包含7512个CDSs (CoDing Sequences), 长度 $\geq 300$ bp的CDSs共有7172个.
  - ▣ DNA双链上6个阅读框中所有长度 $\geq 300$ bp的ORFs都提出来
  - ▣ 共有33527个ORFs, 其中, 7172个CDSs, 26355个非编码的ORFs
- 为研究Set1中ORFs碱基分布模式, 建立理论集Set2:
  - ▣ 基因间序列: 基于7512个注释基因, 从线性染色体上非翻译区提取出所有大于300bp的基因间序列, 共920个
  - ▣ 6个阅读框上所有ORFs的某一个小子集. 共计13281个ORFs
  - ▣ 具体做法可以参考: FEBS Letters 540(2003) 188-194, Hong-Yu Ou.



# Z曲线与原核基因组

- 使用Z曲线分析原核基因组中ORFs的碱基分布
- 数据集Set1:
  - ▣ *S. coelicolor* A3(2)基因组 (GenBank AL645882), 包含7512个CDSs (CoDing Sequences), 长度 $\geq 300$ bp的CDSs共有7172个.
  - ▣ DNA双链上6个阅读框中所有长度 $\geq 300$ bp的ORFs都提出来
  - ▣ 共有33527个ORFs, 其中, 7172个CDSs, 26355个非编码的ORFs
- 为研究Set1中ORFs碱基分布模式, 建立理论集Set2:
  - ▣ 基因间序列: 基于7512个注释基因, 从线性染色体上非翻译区提取出所有大于300bp的基因间序列, 共920个
  - ▣ 6个阅读框上所有ORFs的某一个小子集. 共计13281个ORFs
    - 基于7172个CDSs中的每一个CDSs, 我们要在每个阅读框上找到某一组ORFs. 具体的做法为: 对于7172个CDSs中的每一个在阅读框Forward 0上翻译的CDS, 在对应的编码区中, 对其他5个非编码区阅读框, 都从5'端往3'端找一个ORF ( $\geq 300$ bp). 并且, 对每个非编码阅读框最多只找一个ORF.
    - 例如, 假设一个ORF以ATG开始, 从第二个碱基T开始往下游查找一个在编码区内且长度 $\geq 300$ bp的ORF. 如果找到, 则它被分配给Forward 1这一组.
    - 六个读码框对应ORFs数目为: 7172, 1243, 1215, 638, 2455和558

# Z曲线与原核基因组

## □ 相位特异性Z曲线的方法

- ▣ 设在一个ORF的三个密码子位上, 碱基A,C,G,T出现的频率分别为  $a_i, c_i, g_i$  and  $t_i$ ,  $i=1,2,3$ . 那么:

$$\begin{cases} x_i = (a_i + g_i) - (c_i + t_i), \\ y_i = (a_i + c_i) - (g_i + t_i), \quad i = 1, 2, 3. \\ z_i = (a_i + t_i) - (g_i + c_i), \end{cases}$$

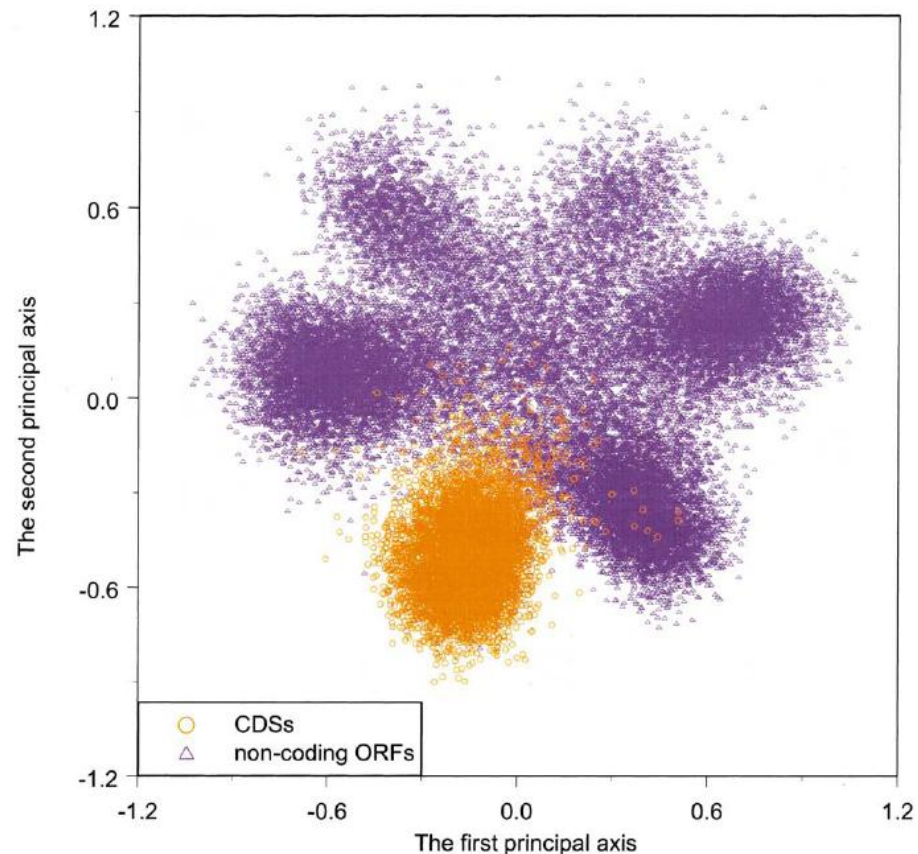
- 每个ORF可以映射为9维空间中的一个点:

$$\begin{cases} u_1 = x_1 - \langle x \rangle, u_2 = y_1 - \langle y \rangle, u_3 = z_1 - \langle z \rangle, \\ u_4 = x_2 - \langle x \rangle, u_5 = y_2 - \langle y \rangle, u_6 = z_2 - \langle z \rangle, \\ u_7 = x_3 - \langle x \rangle, u_8 = y_3 - \langle y \rangle, u_9 = z_3 - \langle z \rangle, \end{cases}$$

- 其中  $\langle x \rangle = (x_1 + x_2 + x_3)/3$ ,  $\langle y \rangle = (y_1 + y_2 + y_3)/3$  and  $\langle z \rangle = (z_1 + z_2 + z_3)/3$

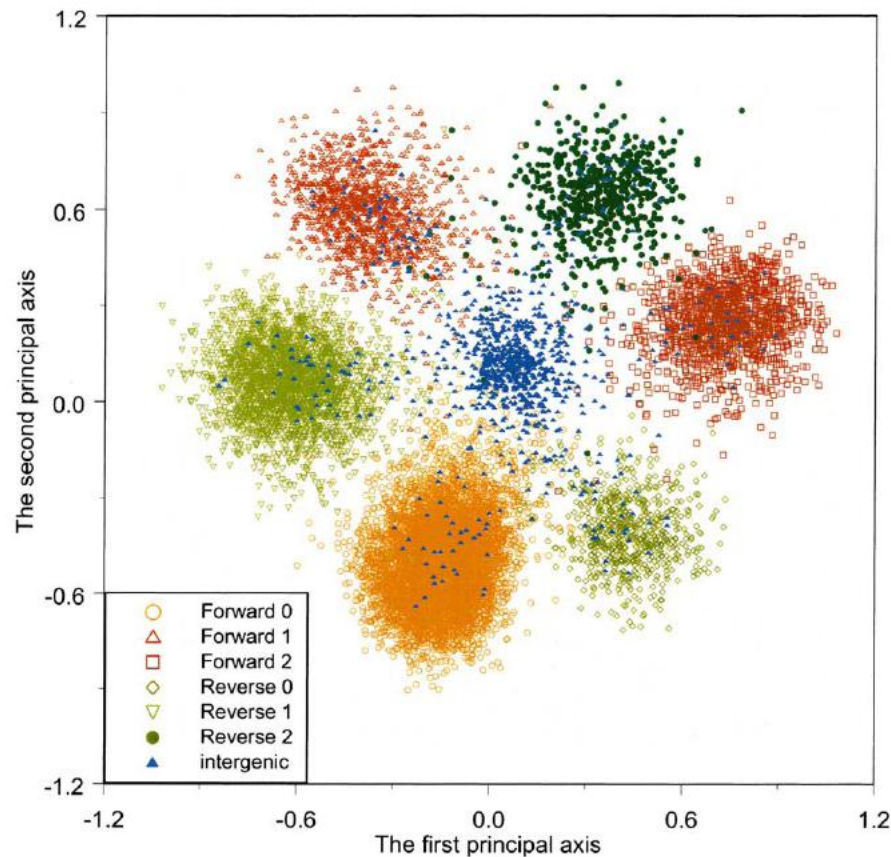
# Z曲线与原核基因组

- 对于数据集set1和set2的每一个ORF,计算9个变量u1-u9,并把它映射到9维空间中一点
- 同时对这两个数据集进行主成分分析(PCA),把它们映射到由第一和第二主成分张成的一个主平面上
- Set1对应的33527个点聚成7类,呈花状结构.其中,最小的一类位于中心,6个花瓣状区域围绕着这个中心区域
  - ▣ 6个花瓣状区域中的一个区域(橙色)对应于7172个CDS
  - ▣ 6个花瓣与中心有一定距离



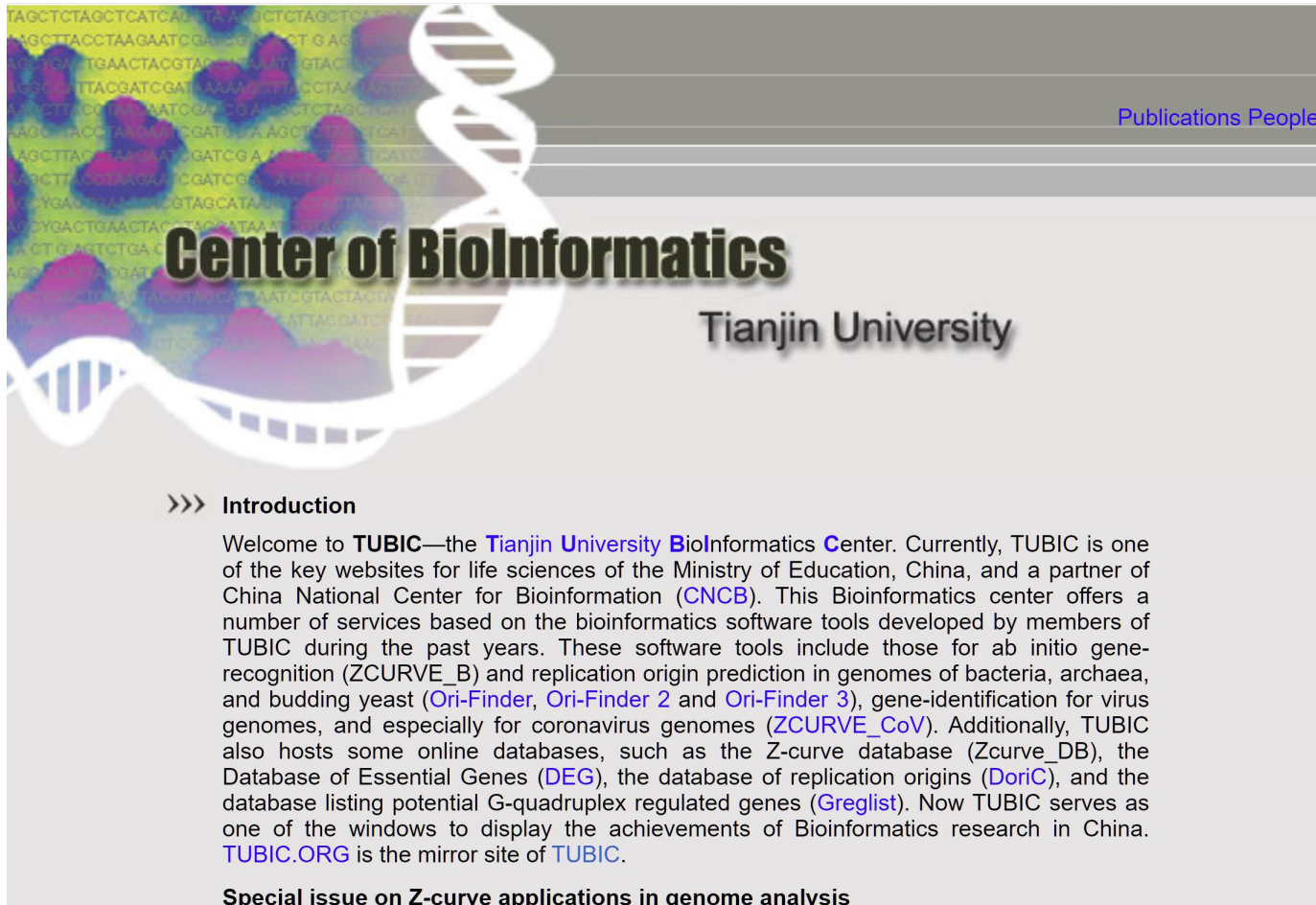
# Z曲线与原核基因组

- **Set2**的结果如图所示
  - 蓝色的中间区域为基因间序列
  - 6类不同相位的**ORFs**分别位于6个花瓣状区域, 并且3个正向阅读框和3个反向阅读框交替出现
  - 橙色花瓣对应于**CDSs**
- 以上结果为高**G+C**含量的原核基因组的**ORF**分布





# 天津大学生物信息学中心

The banner features a background of colorful DNA base pairs (A, T, C, G) and a white DNA double helix structure. The text 'Center of Bioinformatics' is prominently displayed in a large, bold, black font, with 'Tianjin University' in a smaller font below it. In the top right corner, there are links for 'Publications' and 'People'.

Publications People

## Center of Bioinformatics

Tianjin University

>>> Introduction

Welcome to **TUBIC**—the **Tianjin University Bioinformatics Center**. Currently, TUBIC is one of the key websites for life sciences of the Ministry of Education, China, and a partner of China National Center for Bioinformation (**CNCB**). This Bioinformatics center offers a number of services based on the bioinformatics software tools developed by members of TUBIC during the past years. These software tools include those for ab initio gene-recognition (**ZCURVE\_B**) and replication origin prediction in genomes of bacteria, archaea, and budding yeast (**Ori-Finder**, **Ori-Finder 2** and **Ori-Finder 3**), gene-identification for virus genomes, and especially for coronavirus genomes (**ZCURVE\_CoV**). Additionally, TUBIC also hosts some online databases, such as the Z-curve database (**Zcurve\_DB**), the Database of Essential Genes (**DEG**), the database of replication origins (**DoriC**), and the database listing potential G-quadruplex regulated genes (**Greglist**). Now TUBIC serves as one of the windows to display the achievements of Bioinformatics research in China. **TUBIC.ORG** is the mirror site of **TUBIC**.

Special issue on Z-curve applications in genome analysis

<http://tubic.tju.edu.cn/>