＊专题评述＊

# 人与其他生物基因组若干重要问题的生物信息学研究*

张春霆
天津大学生物信息中心，天津 300072

摘要　Z 曲线是表示 DNA 序列的一个等价的三维空间曲线. 通过对 Z 曲线的研究来对基因组序列进行研究是一种几何学的途径. 用这种思路研究了真核和原核基因组中若干重要问题，包括人与高等真核生物基因组的 Isochore 结构，微生物基因组的基因水平转移，古细菌基因组复制起始位点识别，酶母基因组基因识别，细菌与古细菌基因组的 *ab initio* 基因识别，SARS-CoV 基因组基因识别，高 G＋C 含量微生物基因组的结构以及比较基因组学等的研究. 文中综述了天津大学生物信息中心最近 6 年来在上述研究中所取得的进展.

关键词　**基因组　基因组学　生物信息学　几何学　Z 曲线**

天津大学生物信息中心成立于 1998 年初，迄今已有 6 年多的历史. 在这期间，人类基因组计划基本完成，模式生物基因组计划也正如火如荼地进行着，所积累的基因组 DNA 序列的数目成倍地增长，呈“爆炸”之势. 人类与其他生物基因组序列是大自然的伟大作品，它是用一种 4 字母组成的语言写成的“天书”，但是迄今我们人类对这种语言仍然知之甚少. 大自然的奥秘就隐藏在这些厚厚的“天书”之中. 天津大学生物信息中心从一成立就用自己独创的方法参与到解读这部“天书”的国际竞争中来. 非常粗略地说，国内外的生物信息学同行偏重于代数学的途径，而我们则主要采用几何学的方法. 初看之下，一些与基因组学风牛马不相及的几何学名词与概念，如坐标系、空间、投影、曲线、曲率等，在我们的工作中竟然成了分析基因组序列不可或缺的工具. 在这篇综述里，将把我们的主要研究结果进行总结，达到与同行学术交流的目的.

## 1　真核与原核生物基因组基因识别算法研究

我们将基于几何学的基因识别算法研究与基因识别软件的开发列入主要研究方向之一. 下面将介绍这方面的进展.

### 1.1　酿酒酵母基因组基因识别软件 ZCURVE-Y

酿酒酵母(*S. cerevisiae*)是一种单细胞真核生物，是第一个完成测序的真核生物基因组. 我们从其已被实验证实的基因序列出发，提取它们的 Z 曲线的共同数字特征(用 9 个参数描述)，然后用于判别一个待预测的开放读框(ORF)是否为基因[1,2]. 用各种检验方法证实，此算法有 95% 以上的准确度. 在此基础上，我们估计出酿酒酵母基因组基因总数应不多于 5645 个[1]，而不是通常认为的 6000 个[3]. 随后，我们研发出酵母基因组基因识别软件 ZCURVE-Y，并提供网上基因识别服务.

2004-02-10 收稿，2004-06-14 收修改稿
＊ 国家自然科学基金资助项目(批准号：39570187)
E-mail: zzhang@tju.edu.cn

**1.2** 细菌与古细菌基因组 ***ab initio*** 基因识别软件 **ZCURVE**

酵母基因组基因识别算法的发表引起了国际上的广泛重视. 我们尝试将此 9 参数 Z 曲线算法应用于细菌基因组. 在霍乱弧菌(*Vibrio choleræ*)基因组的基因识别中取得了更高的识别精度(98% 以上)[4]. 若从 Z 曲线中提取 18 个参数, 则识别精度可高达 99%以上[5]. 基于 18 个 Z 曲线参数, 我们开发出细菌和古细菌基因识别软件 ZCURVE-C, 并提供网上服务. 它不仅适合任何细菌和古细菌基因组, 而且具有极高的准确率[5]. 但是上述算法的共同缺点是它们要求事先提供待测的基因组中若干已知的基因, 以便训练并提取 Z 曲线参数. 为了克服这一缺点, 随后我们研发出细菌与古细菌 *ab initio* 的基因识别算法与相应软件 ZCURVE[6]. 该算法不需要任何先验的知识, 只要输入待测的基因组的 DNA 序列就可以了. 在此算法中, Z 曲线的提取参数从 18 个增加到 33 个, 准确率又提高了, 但我们用所增加的准确率来换取较低的伪正识别率. 结果, 此算法有 98%以上的准确率, 但有较低的伪正率. 我们把 ZCURVE 版本 1.0 与由美国基因组研究所(TIGR)所开发的、基于 Markov 链的同类 *ab initio* 算法软件 Glimmer 版本 2.02 做了全面的比较[6]. 结果发现, 在识别准确率方面两者旗鼓相当(均为 98%以上), 但 ZCURVE 1.0 比 Glimmer 2.02 具有低得多的伪正率, 尤其对于高 G+C 含量的细菌与古细菌基因组, ZCURVE 1.0 比 Glimmer 2.02 占绝对优势. 在基因起始密码子识别准确率方面, ZCURVE 1.0 明显优于 Glimmer 2.02. 在短基因与水平转移基因识别方面 ZCURVE 1.02 略优于 Glimmer 2.02. 简言之, 即使是 ZCURVE 的早期版本 1.0 也已全面超过了国际著名的 Glimmer 的最新版本 2.02[6]. 不仅如此, ZCURVE 还具有参数少(仅 33 个, 而 Glimmer 所用参数为 1 万个以上), 运行速度快且平稳, 对计算机要求低(一般的微机即可运行), 不仅适用于较大的基因组, 而且对于小基因组也同样适用等一系列优点.

ZCURVE 1.0 的优异表现引起了国际生物信息学与基因组学界的广泛重视. 德国 Göttingen 大学的学者将 ZCURVE 与 Glimmer, CRITICA 和 Orpheus 等国际著名微生物基因组基因识别软件作了全面比较后认为, ZCURVE 1.0 是当前国际上最优秀的微生物基因组基因识别软件之一, 在某些指标上是最好的[7]. 自从 ZCURVE 1.0 提供网上服务以来不到一年, 已有主要来自国外的一千余人次访问或要求提供服务, 先后有 20 余家国内外研究单位与天津大学生物信息中心签定了 ZCURVE 1.0 的机器码转让协议. 发表该软件的论文[6], 已多次被 SCI 刊物引用, 同时该软件也被应用于很多方面. 例如, Egan 等在研究霍乱弧菌时, 用实验学的方法无法验证 GenBank 注释的一个基因, 但该实验结果却与 ZCURVE 对霍乱弧菌基因组的注释相吻合[8]. 与此同时, 发表 ZCURVE 的论文[6]还被评选为 F1000 论文(http://www.facultyof1000.com/).

**1.3** 基因翻译起始位点的识别算法和软件

基因识别研究中的一个重要而困难的问题是基因翻译起始位点的识别. 因为虽然 ATG 等是起始密码子, 但编码区中大部分的 ATG 都是用来编码甲硫氨酸的. 我们发现, Z 曲线在起始密码子附近出现了跳变行为, 而与其上、下游的 Z 曲线行为形成了鲜明的对比. 基于这个发现, 我们提出了新的算法并开发了新的基因翻译起始位点识别软件, 即 GS-Finder[9]. 该软件的识别精度相当高, 优于同类的其他软件, 例如 RBS-Finder. 该软件的另一优点是可以整合到其他基因识别软件中去, 从而提高基因翻译起始位点的识别准确度. 我们已将其整合到 ZCURVE 中, 从而使 ZCURVE 的基因翻译起始位点识别的准确率显著高于 Glimmer[6].

**1.4** 冠状病毒基因组基因识别软件 **ZCURVE-CoV**

冠状病毒(*Coronavirus*)基因组通常很小, 例如引起非典的 SARS 冠状病毒 SARS-CoV 只有不到 30 kb 长. 现有的各种基因识别软件, 如 GeneMark 等应用于 SARS-CoV 基因组往往得到不正确的结果. 为了抗非典的需要, 作者带领几个学生, 很快研发出冠状病毒基因专用基因识别软件 ZCURVE-CoV 1.0 版[10], 该软件尤其适用于 SARS-CoV 基因组, 具有运行速度快、结果准确可靠等优点. 自从 2003 年 6 月份发表并提供网上服务以来, 先后已有 1500 多家研究单位或个人来访或要求提供基因识别服务, 其中包括 WHO 和美国 CDC 等机构. 不久, 我

们又将 ZCURVE-CoV 升级至 2.0 版本. 新版本不仅具有 1.0 版本的一切功能，而且能准确预测 3CL 蛋白酶和 Papain-like 蛋白酶的剪切位点[11]. 这样一来，SARS-CoV 基因组 20 余个基因皆可自动、快速、准确地定位. 运行了 SARS-CoV 2.0，我们将公开发表的 20 余个 SARS-CoV 基因组全部进行了注释. 美国国家生物信息中心(NCBI)对其中某些 SARS-CoV 基因组也进行了注释. 与我们的注释相比，两者基本上是一致的. Van de Hoek 等最近报道了新分离出的 SARS 冠状病毒. 对该病毒基因组的注释使用了 ZCURVE-CoV，并取得很好的结果[12].

### 1.5 基因识别算法的提出、评估与比较

基因识别研究中，算法是关键. 除上述算法以外，我们还提出过基于 Z 曲线的延长-打乱 FFT 算法[13]、长程关联法[14]和终止密码子统计法[15]等. Issac 等将延长-打乱 FFT 算法编成软件并提供网上基因识别服务[16]. 如何评估这些算法的优劣，评估指标的设计至关重要. 我们提出了一系列新指标来评估一个算法的表现[17~20]. 为了直接比较各种基因识别算法的优劣，我们利用人类基因组序列和 EST 数据库建立了一个较大型的外显子、内含子数据库，让待比较的算法逐一来识别它们. 将识别算法的灵敏度与特异度的算术平均值定义为准确度. 按其数值的大小进行排序，这也可以说是一种算法的“比赛”. 国际上先后发表了 Markov 链法、6 核苷酸频率法、密码子使用法和 Z 曲线法等多种基因识别算法. 我们共选取了其中最重要的 19 种算法参加“比赛”. 结果表明：189 个参数和 69 个参数的 Z 曲线算法名列榜首，而 5 阶 Markov 链算法排在第三位[21]. 但是，后者使用了 12288 个参数. 考虑到 ZCURVE 1.0 的优异表现和以上的“比赛”结果，可以肯定地说，在国际上提出的各种基因识别算法中，Z 曲线方法至少是最好的算法之一. Z 曲线方法参数的多少取决于用多少个参数来描述一条三维空间曲线. 国外一组研究人员只用 3 个参数来描述 Z 曲线，进行人类基因识别，并取得了可喜的结果[22].

## 2 人与其他高等真核生物基因组的 Isochore 结构研究

1976 年，意大利学者 Bernardi 及其合作者通过密度梯度超高速离心实验发现人与其他哺乳动物的基因组具有一种“马赛克”式的结构：整个基因组是由一系列 G+C 含量相当均匀的大片段(>300 kb)所组成. 尤为令人惊奇的是，虽然相邻两个大片段的 G+C 含量可以不一样，但从一个片段到另一个片段的过渡，其 G+C 含量是突变的，而不是渐变的. 这种大片段被称之为 Isochore[23]. 其中文意思是：等 G+C 含量的大片段 DNA 序列. 可译为‘等 GC 组成区’. 这里仍沿用 Isochore 这个名字. 进一步的研究表明，不仅是哺乳动物，某些脊椎动物和高等植物基因组也具有 Isochore 结构. 可以说，Isochore 结构是高等真核生物基因组普遍的组织结构形式. 大量的研究证实，Isochore 结构与基因的分布及其表达调控密切相关. 一般认为，阐明人与其他高等真核生物基因组的 Isochore 结构及其生物学含意是理解这些生物基因组结构的关键之一，因而具有重要的科学意义[23].

人类与其他高等真核生物基因组计划的快速进展为 Isochore 结构研究提供了前所未有的机会和条件. 然而事情并不如想像的那么顺利. 目前仍缺乏一种基于序列水平的 Isochore 的定义和识别 Isochore 边界的准确方法. 从一开始人们就赋予 Isochore 两个特点：在其内部 G+C 含量相当均匀；在其边界处 G+C 含量发生了突变. 这样就产生了两个问题：(i)如何定义 G+C 含量的均匀性？(ii)如何识别 G+C 含量的突变点？这两个问题都与基因组中 G+C 含量的计算方法有关. 目前广泛使用的基因组 G+C 含量的计算是沿用窗口方法. 但是，若窗口太大，将抹杀 G+C 含量变化的细节；若窗口太小，将产生很大的统计涨落. 在人类基因组研究中，普遍采用的窗口大小是几十 kb 到几百 kb[24]. 这样的方法是无法识别 Isochore 边界的. 所以，虽然 20 多年前 Isochore 已经被实验所证实，但在人类基因组测序基本完成以后，却无法在基因组序列的水平上找到 Isochore，更不要说确定其边界了[24].

我们发展的几何学方法为解决以上两个问题提

出了新的思路. 我们发明了一种计算 G+C 含量的无窗口方法[25]. 首先将人类基因组序列变换为等价的三维空间曲线——Z 曲线，经过适当的投影和坐标旋转后得到 Z’曲线，后者又称为累积 GC 轮廓图(cumulative GC profile). 我们定义：在基因组某一碱基处的 G+C 含量正比于 Z’曲线在该点切线的斜率. 而在某一窗口中的平均 G+C 含量则正比于此量在该窗口内的定积分. 这样，我们就把生物学中 G+C 含量的概念拓广了，使之在基因组处处都有定义. 如果不从微积分的角度来看，很难理解基因组在某一碱基位置处的 G+C 含量是什么意思. 利用这一无窗口方法，我们计算并研究了人类 1～22 条常染色体和 X，Y 两条性染色体中的 G+C 含量的分布. 我们发现 G+C 含量的变化确是突变式的. 我们从 24 条染色体序列中识别出 56 个 Isochore，从而制定出人类基因组在基因组序列水平上的第一张 Isochore 图谱，并准确确定了全部 Isochore 的边界[26]. 我们发现，即使在 Isochore 内部，其 G+C 含量的变化也是相当大的，但从整个基因组来看还是变化不大的，仍可认为其 G+C 含量是相当均匀的. 我们指出，Isochore 内部 G+C 含量的均匀性是相对的，而不是绝对的[26]. 从研究人类基因组中 G+C 含量的分布曲线，发现人类基因组具有多尺度结构. 小波变换是近年来兴起的新颖的信号处理技术，被称之为数学的显微镜. 从概念上来看，它是研究 Isochore 结构的良好的数学工具. 我们已将其引入到人类基因组的 Isochore 结构研究中来[27]. 我们的研究还证实小鼠和大鼠基因组都具有 Isochore 结构，并在小鼠基因组鉴别出 28 个 Isochore[28]. 此外，还完成了拟南芥(*Arabidopsis thaliana*)基因组 Isochore 结构图的绘制，并发现了一种新类型的 Isochore 结构，称为“着丝粒-isochore”[29]. 关于转座元件(Transposable element，简称“TE”)的研究是一个热点，因为 TE 是真核基因组中的一个重要组成成分. 一般学术界普遍认为在拟南芥基因组中，TE 主要集中于着丝粒附近. 而我们发现，在所有的 5 条染色体中的着丝粒附近，这些富含 TE 的区域却包含了一些只有极少的 TE 的区域，我们将其命名为 TE 沙漠(TE desert). 令人惊奇的是，这些 TE 沙漠的位置与着丝粒-isochore 的位置相吻合[29]. 总之，由于 Isochore 结构与基因组的许多重要功能有关，因此，这些 Isochore 的识别及其边界的确定具有重要的生物学意义.

## 3 细菌基因组水平基因转移研究

细菌的水平基因转移被认为是其进化的普遍方式，细菌通过获得外来基因使之能更好地适应环境甚至产生新的物种. 基因组岛(genomic island)通常包含许多通过水平转移获得的基因. 按其功能不同可分为致病岛、代谢岛、抗抗生素岛等. 某一微生物基因组通过水平转移获得之基因组岛，通常与该生物原有基因组有不同的 G+C 含量、密码子使用与蛋白质氨基酸组成，这就构成了识别基因组岛的基础. 然而实际情况却要复杂得多，因为水平转移来的基因组岛的 G+C 含量可能与宿主基因组的 G+C 含量相当接近，基于窗口的 G+C 含量计算方法由于分辨率低，对这种情况是无能为力的. 而我们提出的 G+C 含量计算的无窗口方法由于拥有高分辨率(其精度可达单核苷酸)，所以对任何情况都适用. 我们用累积 GC 轮廓图成功地识别出许多基因组岛. 例如，蜡状芽孢杆菌(*Bacillus cereus*)与炭疽杆菌(*Bacillus anthracis*)是关系很近的两种细菌，其基因组皆已完成测序. 通过比较这两个基因组的累积 GC 轮廓图，我们从蜡状芽孢杆菌基因组中成功地识别出 3 个基因组岛[30]. 我们还发现，基因组岛的获得有两种模式：一种是外来基因岛单纯地插入原基因组；另一种是在插入的同时伴随着原基因组之基因或基因集团的丢失[30]. 这两种模式被认为在微生物的进化中有着普遍性的意义. 以上论文[30]发表后，该刊物(*Physiological Genomics*)专门刊发了两页的焦点评论(Editorial Focus)文章，对作为工具的累积 GC 轮廓图的重要性以及我们在细菌水平基因转移研究中所做出的贡献给予高度评价[31]. 随后，我们又引入了密码子使用和蛋白质之氨基酸组成偏离指标，并联合累积 GC 轮廓图(Z’曲线)，从谷氨酸棒杆菌(*Corynebacterium glutamicum*)基因组中识别出一个长度为 211 kb 的基因组岛；从创伤弧菌(*Vibrio vulnificus* CMCP6)基因组中分别识别出 3 个基因组岛[32]. 创伤弧菌基因组的基因组岛含有溶血素(hemolysin)基因以及抗多药物(抗生素)泵基因，后者与细菌的抗药性有关. 我们发现这些基因是由水平转移获得的，这对于研究创伤弧菌的致

病性及抗药性有很大帮助. 血红色假单胞菌(*Rhodopseudomonas palustris*)是一种在多种不同环境中广泛存在的细菌，它是一种用于研究代谢途径的模式微生物. 虽然普遍认为该微生物不具有基因组岛，但是我们用累积 GC 轮廓图的方法在它的基因组中发现了 3 个基因组岛[1)]. 其中一个岛携带了编码砷酸盐还原酶 (arsenate reductase) 以及砷排出泵 (arsenical efflux pump) 基因. 砷是一种对大多数微生物有害的半金属. 这两种基因已被证实具有帮助微生物抗砷的功能. 我们发现这两个基因由水平转移获得，这有助于解释基因的水平转移在细菌抗砷中的作用以及该细菌在多种不同环境中生存的原因. 此外，还验证了其他一些微生物基因组中的基因组岛[32]. 研究表明，累积 GC 轮廓图方法不仅适用于基因组岛的识别，还可识别单个水平转移的基因. 这一新方法被认为将成研究微生物基因组水平转移基因的重要的生物信息学工具和标准方法之一[31].

## 4 古细菌基因组复制起始位点的识别

生物进化的三界理论已经普遍被接受，其中古细菌兼具细菌和真核生物的特点. 细菌基因组通常只有单复制起始点，而真核生物基因组普遍具有多复制起始点. 因此，古细菌基因组复制起始位点的研究引起了普遍的关注. 识别细菌与古细菌复制起始点的通用生物信息学方法是 GC-skew 法[33,34]. 但是 GC-skew 法是利用 DNA 游动法从给定基因组序列产生的一种二维曲线，它仅是三维的 Z 曲线的一个特殊情况[35]. 因此，凡是能用 GC-skew 法的地方，Z 曲线也能用. 反之，Z 曲线能解决的，GC-skew 法未必能解决. 以下古细菌基因组复制起始位点的识别就是一些例子. 我们用 Z 曲线方法识别了梅氏甲烷八叠球菌(*Methanosarcina mazei*)基因组的复制起始位点，预测它位于 1564657 bp 与 1556241 bp 之间[36]，而用 GC-skew 方法无效. 詹氏甲烷球菌(*Mathanococcus jannaschii*)是第一个完成测序计划的古细菌，自 1996 年序列公布以来，用了各种理论与实验方法，包括 GC-skew 方法，均不能预测出其复制起始位点. 而我们用 Z 曲线方法确定了其复制起始位点[37]. 我们还用 Z 曲线方法研究盐杆菌(*Halobacterium*)NRC-1 株的基因组，并预测该基因组具有双复制起始位点而且确定了它们的准确位置[38]. 由于古细菌中参与复制的基因非常接近真核生物，人们很早就一直推测古细菌有多复制起始位点，但在该文[38]发表时，所有鉴别出的古细菌复制起始位点全部是单一的. 所以该论文关于双复制起始位点的预测就非常引人注目. 论文发表以后不久，其中的一个复制起始位点的预测就得到实验的证实，实验结果与我们用 Z 方法得到的预测结果准确地吻合[39]. 我们还研究了硫磺矿硫化菌(*Sulfolobus solfataricus*)基因组，并预测该基因组也有多个复制起始位点，并指出了它们的准确位置[38]. 用实验方法对硫磺矿硫化菌基因组复制起始位点识别的竞争非常激烈. 我们的论文[38]发表以后，曾先后有来自丹麦和意大利等几个不同国家的研究小组与我们联系，告诉我们他们的实验正在进行，并希望我们提供关于硫磺矿硫化菌基因组 Z 曲线的进一步信息. 在撰写这篇综述的时候，我们注意到发表在最新一期 *Cell* 杂志上的一篇论文[40]. 在该篇论文中英国和瑞典的研究人员报道，他们已用实验方法在硫磺矿硫化菌基因组中鉴别出两个复制起始位点[40]. 这是历史上用实验方法证实的第一个有多复制起始位点的古细菌. 而这两个用实验方法证实的复制起始位点与我们用 Z 曲线方法预测的位置[38]相吻合. 由此可见，Z 曲线方法是识别细菌与古细菌基因组复制起始位点的强有力的工具.

## 5 高 G+C 含量细菌和古细菌基因组结构研究

基因组之 G+C 含量是决定其组织结构的重要参数. 我们发现[6,41]，G+C 含量在约 56%时是细菌与古细菌基因组结构的一个转折点. G+C 含量大于此值的微生物基因组中 ORF 高度重叠. 通常，描述 ORF 之 Z 曲线的参数组成一个高维空间. 结果，高 G+C 含量的细菌与古细菌基因组的 ORF 在此高维空间形成一个 “6 花瓣” 结构[41]. 天蓝色链霉菌(*Streptomyces coelicolor* A3(2))、绿脓杆菌(*Pseudomonas aeruginosa*)和盐杆菌(*Halobacteri-*

1) Zhang C T, et al. Genomic islands in the Rhodopseudomonas palustris genome. in press

*um*)NRC-1 株(古细菌)都呈现同样的现象，而低 G＋C 含量(小于 56%)的基因组则没有此种现象. 进一步的研究发现，高 G＋C 含量的细菌与古细菌基因组都采用十分相似的密码子使用表，而不管其种、属甚至界是何等的不同[42]. 例如，新月柄杆菌(*Caulobacter crescentus*)、耐放射微球菌(*Deinococcus radiodurans* R1)、盐杆菌 NRC-1 株(古细菌)、结核分枝杆菌(*Mycobacterium tuberculosis*)、苜蓿根瘤菌(*Sinorhizobium meliloti*)、绿脓杆菌和 *Mesorhizobium loti* 基因组均属于高 G＋C 含量的，它们具有十分相似的密码子使用表，而且它们具有共同的基因识别参数[42]. 总之，高 G＋C 含量微生物基因组的组织结构的奥秘还远远没有被揭开，这将是一个重要的研究课题.

## 6 数据库的建立及其他

我们在最近几年里建立了两个国际知名度颇高的数据库. 第一个是基因组 Z 曲线数据库，收集并显示了真核、细菌、古细菌和病毒等 1000 多个基因组的三维 Z 曲线[35]. 为方便研究人员使用 Z 曲线，我们还编写了 Z 曲线绘图软件 Zplotter. 该软件可在网上对用户输入的 DNA 序列绘制出 Z 曲线，包括累积 GC 轮廓图，同时可有放大缩小功能，并具有计算并输出 Z 曲线坐标等功能[35]. 自 2003 年 3 月份提供网上查询、显示和分析服务以来，已有 2000 多名国内外研究人员上网访问，使用该数据库. 发表此数据库的论文[35]也被英国现代生物出版集团评为 F1000 论文. 对上千个基因组 Z 曲线的分析导致了一个新生物统计量的提出，我们称之为基因组序指数，它综合反映了基因组的组成特征[43]. 第二个数据库为微生物必需基因数据库 DEG (database of essential genes)，收集了被实验证实的 2000 余个微生物必需基因序列与功能注释，并可对用户所提供的待分析基因进行数据库的 Blast 搜索服务[44]. 该数据库在药物设计中有广泛的应有. 例如，许多必需基因的产物是抗菌药物设计中很好的靶点. 数据库 DEG 自从 2004 年 1 月 1 日发表于 Nucleic Acids Research 的数据库专辑以来，已有上千名研究人员访问并要求提供搜索服务. Humana 出版社发行了一套比较著名的系列丛书，名为 Methods in Molecular Biology. 我们应邀在其中的一个分册，Gene essentiality at genome scale 中撰写一个章节，主要介绍如何利用 DEG 进行基因关键性的分析[1)].

除了上述研究以外，还用 Z 曲线方法研究了蛋白质的亚细胞定位，或用类似方法研究了球蛋白质结构的分类以及肽链的可折叠性等问题，详见有关论文[45～50].

## 7 网上服务信息

我们从 2000 年开始建立了天津大学生物信息中心网站. 网站定位在向国内外的生物信息学研究人员免费提供独立研发的各种软件服务. 该网站开通 3 年多来，已经成为向国外同行展示我国生物信息学研究成果的窗口之一. 天津大学生物信息中心的简称‘TUBIC’，似乎已经成了生物信息学界的一个知名品牌. TUBIC 网站服务信息详情可见表 1.

**表 1 天津大学生物信息中心网站服务项目信息**

| 项目 | 版本 | 网址 | 注释 |
|---|---|---|---|
| TUBIC | —— | http://tubic.tju.edu.cn | 天津大学生物信息中心网站主页 |
| ZCURVE | 1.02 | http://tubic.tju.edu.cn/Zcurve_B | 细菌和古细菌全基因组 *ab initio* 基因识别[6] |
| Zcurve_C | 1.0 | http://tubic.tju.edu.cn/Zcurve_C | 细菌和古细菌基因组中单个基因识别[5] |
| Zcurve_Y | 1.0 | http://tubic.tju.edu.cn/Zcurve_Y | 酿酒酵母基因组基因识别[1] |
| Zcurve_CoV | 2.0 | http://tubic.tju.edu.cn/sars | 冠状病毒(含 SARS-CoV)基因组基因识别[10,11] |
| GS-Finder | 1.0 | http://tubic.tju.edu.cn/GS-Finder | 细菌和古细菌基因起始密码子位点识别[9] |
| Zcurve_DB | 1.0 | http://tubic.tju.edu.cn/zcurve | 1000 余种生物基因组的 Z 曲线数据库[35] |
| DEG | 1.1 | http://tubic.tju.edu.cn/deg | 细菌和古细菌必需基因数据库[44] |
| HighGC | 1.0 | http://tubic.tju.edu.cn/highGC | 高 G＋C 含量的细菌与古细菌基因组研究[41] |

1) Zhang C T, et al. Gene essentiality analysis based on DEG, a database of essential genes. In: Gene Essentiality at Genome Scale: Protocols and Bioinformatics. Series: Methods in Molecular Biology. in press

## 8 总结与展望

天津大学生物信息中心实际上是一个很小的研究小组，由一名教师(即作者本人)再加上几个流动性很大的学生所组成，花钱也不多. 我们 6 年多的实践证明，将基因组序列一一对应地变换为三维空间曲线(Z 曲线)，再通过对 Z 曲线的研究来研究基因组，这样的思路是切实可行的. 例如，在基因识别问题中，传统的方法是分别计算编码和非编码序列中大量的概率和条件概率，通过对这些概率的比较来区别它们. 而我们则通过对编码与非编码序列的 Z 曲线的比较来区别它们，方法既简单，效果又好. 原则上说，基因组中的许多问题都可以通过这种途径加以解决. 6 年来先后发表 SCI 论文 40 余篇，本文介绍了我们在这方面研究所取得的进展. 这种独树一帜别开生面的研究思路已经得到国内外学术界的普遍好评和认可，越来越多的同行，主要是国外同行，加入到对 Z 曲线研究的行列中来. 可以预期，用几何学方法研究基因组将会有一个广阔的发展空间. 我们的实践还证明，生物信息学是一门花钱少、见效快，非常适合中国国情的学科. 只要加以适当的组织和引导，经过不懈的努力，在不太长的时间内，完全有可能使生物信息学成为我国的优势学科之一.

## 参 考 文 献


1 Zhang C T, et al. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. Nucleic Acids Res, 2000, 28: 2804

2 Zhang C T, et al. Using a Euclid distance discriminant method to find protein coding genes in the yeast genome. Comput Chem, 2002, 26: 195

3 Goffeau A, et al. Life with 6000 genes. Science, 1996, 274: 546, 563

4 Wang J, et al. Identification of protein-coding genes in the genome of Vibrio cholerae with more than 98% accuracy using occurrence frequencies of single nucleotides. Eur J Biochem, 2001, 268: 4261

5 Chen L L, et al. Gene recognition from questionable ORFs in bacterial and archaeal genomes. J Biomol Struct Dyn, 2003, 21: 99

6 Guo F B, et al. ZCURVE: A new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acids Res, 2003, 31: 1780

7 Tech M, et al. YACOP: Enhanced gene prediction obtained by a combination of existing methods. In Silico Biol, 2003, 3: 0037

8 Egan E S, et al. Distinct replication requirements for the two Vibrio cholerae chromosomes. Cell, 2003, 114: 521

9 Ou H Y, et al. GS-Finder: A program to find bacterial gene start sites with a self-training method. Int J Biochem Cell Biol, 2004, 36: 535

10 Chen L L, et al. ZCURVE-CoV: A new system to recognize protein coding genes in coronavirus genomes, and its applications in analyzing SARS-CoV genomes. Biochem Biophys Res Commun, 2003, 307: 382

11 Gao F, et al. Prediction of proteinase cleavage sites in polyproteins of coronaviruses and its applications in analyzing SARS-CoV genomes. FEBS Lett, 2003, 553: 451

12 Van Der Hoek L, et al. Identification of a new human coronavirus. Nature Med, 2004, 10: 368

13 Yan M, et al. A new fourier transform approach for protein coding measure based on the format of the Z curve. Bioinformatics, 1998, 14: 685

14 Zhang C T, et al. A novel approach to distinguish between intron-containing and intronless genes based on the format of Z curves. J Theor Biol, 1998, 192: 467

15 Wang Y, et al. Recognizing shorter coding regions of human genes based on the statistics of stop codons. Biopolymers, 2002, 63: 207

16 Issac B, et al. Locating probable genes using Fourier transform approach. Bioinformatics, 2002, 18: 196

17 Zhang C T, et al. A graphic approach to evaluate algorithms of secondary structure prediction. J Biomol Struct Dyn, 2000, 17: 829

18 Zhang C T, et al. A refined accuracy index to evaluate algorithms of protein secondary structure prediction. Proteins, 2001, 43: 520

19 Zhang C T, et al. Evaluation of gene-finding algorithms by a content-balancing accuracy index. J Biomol Struct Dyn, 2002, 19: 1045

20 Zhang C T, et al. $Q^9$, a content-balancing accuracy index to evaluate algorithms of protein secondary structure prediction. Int J Biochem Cell Biol, 2003, 35: 1256

21 Gao F, et al. Comparison of various algorithms for recognizing short coding sequences of human genes. Bioinformatics, 2004, 20: 673

22 Wu Y H, et al. Classification of short human exons and introns based on statistical features. Physical review E, 2003, 67: 061916

23 Bernardi G. The human genome: Organization and evolutionary history. Annu Rev Genet, 1995, 29: 445

24 Lander E S, et al. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 860

25 Zhang C T, et al. A novel method to calculate the G+C content of genomic DNA sequences. J Biomol Struct Dyn, 2001, 19: 333

26 Zhang C T, et al. An isochore map of the human genome based on the Z curve method. Gene, 2003, 317: 127

27 Wen S Y, et al. Identification of isochore boundaries in the human

genome using the technique of wavelet multiresolution analysis. Biochem Biophys Res Commun, 2003, 311: 215

28 Zhang C T, et al. Isochore structures in the mouse genome. Genomics, 2004, 83: 384

29 Zhang R, et al. Isochore structures in the genome of the plant Arabidopsis thaliana. J Mol Evol, 2004, 59: 227

30 Zhang R, et al. Identification of genomic islands in the genome of Bacillus cereus by comparative analysis with Bacillus anthracis. Physiol Genomics, 2003, 16: 19

31 Charkowski A O. Making sense of an alphabet soup: The use of a new bioinformatics tool for identification of novel gene islands. Focus on "identification of genomic islands in the genome of Bacillus cereus by comparative analysis with Bacillus anthracis". Physiol Genomics, 2004, 16: 180

32 Zhang R, et al. A systematic method to identify genomic islands and its applications in analyzing the genomes of Corynebacterium glutamicum and Vibrio vulnificus CMCP6 chromosome I. Bioinformatics, 2004, 20: 612

33 Grigoriev M. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res, 1998, 26: 2286

34 Lobry J R A. simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. Biochimie, 1996, 78: 323

35 Zhang C T, et al. The Z curve database: A graphic representation of genome sequences. Bioinformatics, 2003, 19: 593

36 Zhang R, et al. Single replication origin of the archaeon Methanosarcina mazei revealed by the Z curve method. Biochem Biophys Res Commun, 2002, 297: 396

37 Zhang R, et al. Identification of replication origins in the genome of the methanogenic archaeon, Methanocaldococcus jannaschi. Extremophiles, 2004, 8: 253

38 Zhang R, et al. Multiple replication origins of the archaeon Halobacterium species NRC-1. Biochem Biophys Res Commun, 2003, 302: 728

39 Berquist B R, et al. An archaeal chromosomal autonomously replicating sequence element from an extreme halophile, Halobacterium sp. strain NRC-1. J Bacteriol, 2003, 185: 5959

40 Robinson N P, et al. Identification of two origins of replication in the single chromosome of the archaeon Sulfolobus solfataricus. Cell, 2004, 116: 25

41 Ou H Y, et al. Analysis of nucleotide distribution in the genome of Streptomyces coelicolor A3(2) using the Z curve method. FEBS Lett, 2003, 540: 188

42 Chen L L, et al. Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. Biochem Biophys Res Commun, 2003, 306: 310

43 Zhang C T, et al. A nucleotide composition constraint of genome sequences. Comput Biol Chem, 2004, 28: 149

44 Zhang R, et al. DEG: A database of essential genes. Nucleic Acids Res, 2004, 32 Database issue: D271

45 Feng Z P, et al. A graphic representation of protein sequence and predicting the subcellular locations of prokaryotic proteins. Int J Biochem Cell Biol, 2002, 34: 298

46 Zhang C T, et al. A quadratic discriminant analysis of protein structure classification based on the Helix/Strand content. J Theor Biol, 1999, 201: 189

47 Zhang C T, et al. A new criterion to classify globular proteins based on their secondary structure contents. Bioinformatics, 1998, 14: 857

48 Zhang C T, et al. A new quantitative criterion to distinguish between alpha/beta and alpha+beta proteins (domains). FEBS Lett, 1998, 440: 153

49 Zhang C T, et al. S curve, a graphic representation of protein secondary structure sequence and its applications. Biopolymers, 2000, 53: 539

50 Zhang C T, et al. Skewed distribution of protein secondary structure contents over the conformational triangle. Protein Eng, 1999, 12: 807