基因组信息、密码进化、折叠动力学和熵产生
——理论生物学的几个基本问题

罗辽复

内蒙古大学物理学院,呼和浩特 010021

摘要 数千个基因组的测序完成,海量资料数据呼唤着生命科学的理性和实证性更完美地结合。基因组信息学是理论生物学的龙头。本文遵循生命信息运行的密码—序列—结构—功能的主线,论述了生命科学的理性化途径,讨论了若干基因组生物学的基本问题。着重总结了理论生物学领域的部分工作,包括:① 基因组序列识别码的形成和构建,一组描述碱基分布和碱基关联的统计量集合可作为基因组的识别码;② 基因组的进化方向和最大信息原理,提出基因组序列的编码信息量在进化中随时间增长的规律,指出 DNA 序列局部片段遵循最大信息原理;③ 遗传密码的适应性进化,证明普适标准密码表是突变危险性的适应性极小码,论证了遗传密码既具有稳定性,又具有可进化性;④ 基于量子跃迁的蛋白质折叠动力学,用量子跃迁的观点和理论研究如何处理从序列到结构,导出蛋白质折叠速率公式,解释了现有各种速率实验;⑤ 细胞开关相变和熵产生,研究了最简单生命噬菌体的溶原态/裂解态转变动力学,得到了此类细胞开关中熵产生的特异性规律。

关键词 基因组信息;进化;遗传密码;蛋白质折叠;熵产生

中图分类号 Q61 **文献标识码** A **文章编号** 1000-7857(2010)15-0106-06

Genome Information, Code Evolution, Folding Dynamics and Entropy Production: Several Fundamental Problems in Theoretical Biology

LUO Liaofu

Department of Physics, Inner Mongolia University, Hohhot 010021, China

Abstract The rationalization of biology is discussed in an information-centered context. The studies in the field of theoretical biology are reviewed which include the following topics. (1) The formation and construction of the genome recognition code. A set of statistical quantities describing base composition and base correlation can serve as the genome recognition code. (2) The genome evolution direction and the maximum information principle. It is proposed that the quanlity of the function-coding information of a genome grows with time in the course of evolution. It is proposed also that the Shannon information quantity of the local DNA segments obeys the maximum information principle. (3) The adaptive evolution of the genetic code. It is proved that the prevalent standard amino acid code is a mutational deterioration-minimal code of an adaptive evolution and it is demonstrated that the genetic code satisfies both the principles of robust stability and evolvability. (4) The protein folding dynamics based on quantum transition. The protein folding rate formula is deduced based on quantum transition between torsion states by using conformation dynamics. All theoretical results (including the rate and its dependence on chain length, inertia moment and temperature, etc.) are consistent with the updated experimental data. (5) The cell switch and entropy production. The dynamics of lysogenic/lytic transition of lambda phage – the simplest form of the life – is studied from a set of differential equations. The entropy production rate of the typical cell switch is calculated. It is proved that the dynamics of this particular system obeys some topological theorem and thus the obtained results are of general nature.

Keywords genome information; evolution; genetic code; protein folding; entropy production

---

收稿日期: 2009-11-23;**修回日期**:2010-07-17

基金项目: 国家自然科学基金项目(90103030,90403010);内蒙古自治区科学技术特别贡献奖励基金(2008)

作者简介: 罗辽复(中国科协所属全国学会个人会员登记号:S250700696M),教授,研究方向为理论生物物理学,电子信箱 lolfcm@mail.imu.edu.cn

0 信息是生命的精髓

实证性和理性的结合是近代科学的特点和优点,这种结合在物理学中表现得最为完美和富有成果,并正在渗透到自然科学的其他部门。传统的生物学都是实验的,"正在建立的研究新模式是:由于全部基因将被知晓,存储在电子数据库中,生物学研究的出发点将是理论的。一个科学家将从理论假说出发,然后转向实验,去追随和检验这些假说。"[1] 因此,生命科学正在从传统的单一的实验方法发展到与逻辑的理性的方法相结合,从一门实验科学提升到定量的、理性的水平上来。这种理性化的努力从分子生物学诞生之日起,探索一波又一波地进行着[2-6]。

过去的自然科学基本上聚焦于物质和能量两个基本范畴。不同于物质和能量,与其相平行的"信息"是自然界的第三个基本范畴。生命系统的特征就在于它包含的大量信息,生命始于信息,信息是生命的精髓。因此,理论生物学研究应以生命信息运行中的基本问题为中心。最近 15 年发展起来的生物信息学是生物学理性化潮流的一部分,它已显示出巨大的冲击力,但作者认为现时很多的生物信息学工作可能过于"工程化"。本文将挑选若干生命信息运行中的理论生物学基本问题分别进行介绍评述和展望。

1 基因组序列识别码的形成和构建

密码—序列—结构—功能是生命信息运行的固有逻辑。密码指信息的编码规则,序列是信息的载体,一个大分子序列包含极多坐标,只有在给定结构中才具有生物活性,才能与其他分子实现正确的物理作用,从而保证一定生物化学过程的进行,实现某种确定的生物功能;而结构是由序列决定的。在这个逻辑路线中,给定了编码规则后,序列就是信息的运行和表达过程的起点。如何识别基因组的序列特征是否存在基因组的序列识别码?

1991 年,罗辽复等把二阶信息冗余 D_2 推广为互信息的一般表示:[7]

$$D_{k+2}=2H+\sum_{ij}p_{i(k)j}\mathrm{lb}p_{i(k)j}=\sum_{ij}p_{i(k)j}\mathrm{lb}\frac{p_{i(k)j}}{p_ip_j}\qquad k=0,\ 1,\ 2,\ \cdots$$

这里,$p_{i(k)j}$ 为相距 k 的一对碱基 i 和 j 的联合概率,$H=-\sum_i p_i\mathrm{lb}p_i$ 为信息熵。提出用一组描述碱基分布和碱基关联的统计量集合$(D_1,D_2,\cdots,D_s)$描述一个基因的特性,并构建了进化关系。D_1 为一阶信息冗余,描写遗传语言词汇(核苷)组成的非随机性,$D_{k>1}$ 为(推后)二阶信息冗余,描写各种碱基关联或遗传语言的语法结构。同时,证明了有限长度的无规或独立序列的 D_1 服从自由度为 3 的 χ^2 分布,$D_{k>1}$ 服从自由度为 9 的 χ^2 分布,导出了一个长度 N 的 DNA 序列的信息参数 D_1、D_2 等的涨落限均反比于序列长 N。对基因序列的统计研究结果表明,大部分序列的 D_{k+1}–k 图形具有共同特征,D_2 最大,近程关联为主,而 k 大的那些 D_k 常常落到涨落限以下。但对于整条染色体或长度大于 10^4 的序列来说,涨落就很小了。另外,考虑到

$$D_2\cong\frac{1}{\ln 2}\sum_{ij}\frac{(p_{ij}-p_ip_j)^2}{p_ip_j}$$

求和中有 16 项,代表 16 种碱基关联(D_3 等有类似表式)。1996–1998 年,本研究组研究了 DNA 序列中特定类型的碱基关联及其涨落特性[8-9]。

此后,Grosse 等(2000)[10]、Yu 等(2001)[11]、Otu 等(2003)[12]从不同角度用 $\{D_k\}$ 识别序列和讨论序列进化。特别是 2008 年,Bauer、Schuster、Sayood 用$\{D_k\}$分析染色体和基因组,强调指出平均互信息(Average Mutual Information,AMI),即我们定义的信息冗余 D_k 可以作为基因组的特征信号[13]。

一个有兴趣的问题是,基因的 D_k 常在涨落限下,而基因间的协同作用使基因组的 D_k 携带信息,使这些 D_k 变成基因组的特征信号。特征识别码是如何形成的?统计分析表明,当基因组中所考虑的序列长度大于 N_c (N_c~2×10^5),D_k 的信息就超过了涨落限[14]。故 N_c 是随机涨落和功能有序性的转变点,这种转变是"生命发生于微观宏观之间"[15]的一个例证。

除了用$\{D_k\}$,还可用 k–mer 频数识别 DNA 序列和讨论序列进化。k–mer 频数分布表达了基因组中序列组织的规律性,国内外有许多这方面的工作,如 Reinhart 等[16],Hao 等[17],Sims 等[18],罗辽复等对 k–mer 的距离保守性[19]和频数求和规则[14]等问题的研究也获得了一些有意义的结果。

遗传信息从 DNA 传递到 RNA 序列后,RNA 结构对信息的进一步运行有重要作用。本课题组研究了 mRNA 的折叠方式和折叠能量,证明基因组内不同基因的折叠能差别统计上远小于基因组间的差别[20];最近还证明了对于 5′–和 3′–非翻译区以及内含子的 RNA 折叠能也有类似性质,因此表明可能存在基于 RNA 序列的 RNA 折叠码。

过去认为遗传信息基本上全部储存在 DNA 序列之中,新近发现 DNA 甲基化组蛋白修饰和核小体定位对真核基因的表达调控有重要作用[21],这些信息是否构成了独立于 DNA 序列的另一类信息源,它们如何协同地在基因表达和基因组识别中发挥作用,是一个有兴趣的理论问题。

2 基因组的进化方向和最大信息原理

达尔文"物竞天择,适者生存"的思想本身就说明进化是有方向的。他着重解释了"性状分歧"和进化树的形成,"生殖率如此之高以致引起生存竞争,因而导致自然选择,并引起性状分歧和较少改进的类型的灭绝";而进化的树状图示更深刻地表明了进化的方向性。这种方向性如何从分子生物学的水平和基因组学的角度予以说明?不少研究者试图从物种的基因组尺度阐明进化规律,但由于"C 值佯谬"(基因组的进化复杂性与其大小没有关系),这些努力均告失败[22]。综合现有实验资料,作者强调信息量在物种的功能革新和扩展中的作用,提出了下述关于基因组进化方向的信息扩增律[23-24]。

在 DNA、RNA、蛋白质的相互作用下,通过序列复制和编码方式扩展,以及基因在基因组间转移,基因组序列的编码

信息量 I_c 在进化中随时间增长($dI_c/dt \geq 0$)。

与能量不同,信息不具守恒性。信息的扩增是生物学的基本规律,它在生物学中的意义如同物理学的能量转换和守恒。薛定谔的微型密码观念对 10 年后分子生物学的诞生极其重要,但它忽略了一个重要方面:没有考虑遗传信息是如何积累扩增和进化的。

令编码一定功能的序列由符号串 $A_1, A_2, \cdots, A_n$ 表示,其中 A_i 可取 S_i 个可能值,上述扩增律中编码信息量定义为 $I_c = \mathrm{lb}\prod_i S_i$。编码信息量包含两部分,一是由编码蛋白质的 DNA 序列所贡献,称为 p–编码信息量;二是由编码其他功能(包括基因调控)的基因组序列所贡献,称为 n–编码信息量。这个规律与现有实验资料相符,因为各门或纲基因组最小值从原核生物到真核哺乳动物是依次增加的,这粗略反映了编码蛋白质的信息量随物种遗传复杂性增加;再考虑基因调节机制随着基因组功能的复杂性变化(n–编码信息量的贡献增长),则编码信息量随物种进化的图像就会很清晰。罗辽复等分析了基因组进化中的一些重大事件,如全基因组复制和基因丢失,寒武纪大爆发等几次适应性辐射,以及灵长类到人类基因组的进化等,均未发现与编码信息量扩增律矛盾的情况。

这个规律的着眼点是基因组随时间变化,物种的复杂性是通过进化中出台的时间顺序而被考虑的。dt 的精确度决定于自然选择对遗传过程起作用的最小时间间隔(若干世代)。

编码信息量的增长速度可作为物种进化速率的标度。进化与环境变化密切相关,在环境剧烈变动中,大量物种灭绝,是由于老物种 I_c 的增长速度赶不上环境变化,让位于那些具有更适应于新环境的功能需求的基因的新物种。

细菌在宿主中的寄生导致它失去一部分功能,从而减少了基因组的编码信息量。这是一种退化现象,不属于本定律描述的范围,本定律只适用于自由生活基因组。

2000 年,Kauffman 提出了生物系统中可能存在热力学第四定律的问题[25],上述基因组信息扩增律与此思想一致,但更具本质意义的是对于所研究的系统和规律的形式做出了具体规定。

如果基因组信息扩增律是正确的,就可以研究每一基因组的编码信息量的演化,并与其他物种的基因组比较,从总体上把握各基因组的垃圾 DNA 情况,估算其中可能包含的编码关系,搞清楚这部分“暗信息”的意义。这也将有助于全面认识真核基因调节网络,从分子水平上解决生命活动如何在时间轴上展开这一基本问题。

上述扩增律中编码信息量定义为编码状态数的对数(Boltzmann 信息量),与编码信息量相关的是各种编码状态出现频数如何分布、如何变化的问题。令此频数分布为 $\{p_i\}$,定义 Shannon 信息量 $I_S = -\sum_i p_i \mathrm{lb} p_i$。频数分布变化受两个互相矛盾互相补充的因素制约:一是碱基的随机突变,二是功能约束造成的碱基保守性,以及功能本身的进化(功能扩展或功能丢失)。因此,状态频数的变化表现出复杂的进化关系。但由于功能约束的进化相对缓慢,而碱基突变是较快发生的事件,DNA 序列还是显示出一定功能约束下碱基突变的随机性。最大信息原理指出,非平衡系统的概率分布可由服从一定约束下的信息量取极大值导出。本小组首次将此原理用于核酸系统,证明了在一组功能约束和进化约束条件下,DNA 序列碱基频数分布的 Shannon 信息量是最大的[26–27]。因此,基因组的进化一方面整体上是编码信息量扩增的,是 Lyapounov 不稳定的;另一方面在给定的编码方式下,DNA 序列局部片段是 Shannon 信息量极大化的。这个情况类似人类语言的进化。

序列的多样性等于 Shannon 信息量与序列长的积,在把最大信息原理看作研究 DNA 序列碱基分布的基本原理的基础上,本研究组提出了识别序列模式的 IDQD 算法(多样性增量二次判别法),在生物信息学中获得了成功应用[28–30]。

3 遗传密码的适应性进化

遗传密码(氨基酸码)具有高度普适性,而非标准码的出现又说明它仍在进化着,必须从密码表构成的基本角度,才能理解这种普适性和可进化性。1988—2002 年,罗辽复等建立了遗传密码的构成和进化的统一理论,包含如下 5 个要点[31–36]。

1) 提出了包含碱基突变频率和由它产生的氨基酸取代而导致的选择性死亡两个因素的突变危险性概念,证明密码的简并规则(冗余性分布)具有突变危险性极小的性质。突变危险性(Mutational Deterioration,MD)是指码字(密码子)的突变导致氨基酸误编码的有害效应。这种有害效应必然为进化中的自然选择所消除,因而观察到密码子的冗余性分布遵守 MD 极小原理。即,一个多重态在密码表上的分布,与假想分布相比,具有 MD 极小性。假定,第一,对于每一假想的密码子多重态,可定义突变危险性系数 MD,代表这个多重态受到突变干扰的频率和引起的危险性。第二,突变危险性系数分为 3 类:非同义的单碱基转换突变系数 u,单碱基颠换突变系数 v 和密码子第三位上附加的摆动突变系数 w_u 或 w_v。假想的密码子多重态的总 MD 值等于其包含各个密码子 3 个碱基的 u、v、w_u、w_v 值之和。根据以上两个假设,笔者证明了现有密码表上各个多重态(包括终止密码)的简并规则都是满足 MD 极小化原则的,只要参数 u、v、w_u、w_v 的相对大小满足一组和实验资料一致的相当宽的约束条件($u>2v$, $w_v>u-v$, $w_u-w_v>u+v$)即可。由此解释了密码表的冗余性分布,说明了这种分布的鲁棒性。

2) 导出了密码表上亲疏水氨基酸的畴状分布。在多重态局域突变危险性导出简并规则的基础上,进一步从密码表的全局突变危险性(Global Mutational Deterioration,GMD)极小化导出了密码表的亲水–疏水畴。同时,引入 4 种核苷酸的阴阳对偶性(UCGA 间既具有内部对称性又具有系统性质差异的一种表示),从阴阳对偶性更深刻地说明密码表中亲水氨基酸和疏水氨基酸的畴状分布的必然存在和鲁棒性。

3) 提出密码表的全局突变危险性作为进化(不)适合度(non–fitness)的度量,并导出 GMD 极小的表。全局突变危险

性依赖于两个因素，一是用参数 u、v、w_u、w_v 表达的密码子的遗传距离 f_{ij}，二是所编码氨基酸的距离 $D_{\alpha\beta}$，

$$\mathrm{GMD}(U)=\sum_{i\neq j\alpha\beta}U_{i\alpha}U_{j\beta}f_{ij}D_{\alpha\beta}$$

GMD 的极小化相应于 Wright 适应性地形上的极大化。同时，GMD 也是一个误差函数，反映了碱基突变误差和翻译误差。与国外的误差极小理论[24]不同，GMD 的极小化具有独立的进化含义；并由于此理论中两种误差可分别处理，突变误差已由 u、v、w_u、w_v 表达，从简并规则定出，因此 GMD 极小化可以更方便地进行且得到更明确的结果。笔者发现，GMD 极小表出现在标准密码表下方约 10%的位置上。

4) 证明普适标准密码表是适应性 GMD 极小码。适应性指适应于氨基酸进化早期的条件，利用给定约束下的适合度函数，GMD 极小化可以证明，标准密码表是 20 种氨基酸的简并度约束，以及与氨基酸合成途径有关的早期编码约束下的最优编码。与国外一些理论[37]不同，文献[31]~[36]中的标准密码表是进化可及的。另外有意思的是，标准密码表可能是在进化早期现存生命的最近共同前体出现后的很短时间内爆炸式地形成的。

5) 密码表的进化和反常密码的产生。与 tRNA 的易变性相伴随，氨基酸数量可能扩展(如 21 号、22 号氨基酸)，终止密码和氨基酸的简并度可能变化，这些因素都将导致约束条件的变化，从而引发适合度函数极值变化和反常密码的产生。罗辽复等提出利用各密码子多重态的 MD 极小值估算由于约束条件变化引起的 GMD 变化的方法，证明了现有 30 种非标准密码与标准密码相比，基本上都是 GMD 减小的；它们是适应于变化的约束条件的最佳码。具体说来，反常密码有 5 种类型：① 终止密码编码氨基酸；② 有意义密码子的无意义化；③ 通过密码子俘获的氨基酸编码的改变；④ 通过不确定中间态的氨基酸编码的改变[38-39]；⑤ 从有意义密码子转变为终止密码再转变为有意义密码子的两步过程。除了 3 例第②类反常外，GMD 都是减小的；这 3 个例外可能与有意义密码子无意义化的计算尚存一定问题有关[36]。从应用角度看，弄清了密码的进化规律，就有可能人工改变密码，为基因工程和蛋白质工程开辟新的途径。

总体上看，遗传密码的进化包含两个方面，一是给定约束(氨基酸数量和简并度)下适合度函数 GMD 的极小化；二是约束条件的变化导致 GMD 极值的变化。因此，遗传密码既具有稳定性，又具有可进化性。这些情况类似于人类语言的演化。相信对于氨基酸码以外的其他密码的形成和进化，这个规律同样适用。

4 基于量子跃迁的蛋白质折叠

生命信息运行中的一个关键步骤是从序列到结构，而结构的形成和变化离不开物理原理。笔者工作发现，蛋白质折叠可能和量子跃迁密切相关。

分子生物学系统的基本微观变数是分子构象和前沿电子。这里的构象主要指扭角，扭角是大分子形状中最易改变的部分。扭转振动的频率在微波范围，可以证明系统的信息量主要凝聚在这个波段。故生物大分子的活性除了前沿电子坐标外，主要决定于扭角坐标$\{\vartheta\}$。基于构象运动和电子运动的相互作用是生物活性微观基础的认识，笔者提出了构象动力学理论[40]。它不同于主要讨论电子运动的量子生物化学，也不同于经典力学基础上的分子动力学，适宜于从量子观点研究构象变化。

扭转势能 U 作为 θ 的函数，一般有几个极小值，对应于几个构象 k；据量子力学理论，在这种对称势函数中状态不可能是局域的。构象动力学证明了只要势函数有微小不对称，便可定义局域的构象态 ks。一般地，构象电子系统的波函数在绝热近似下表示为

$$M_{ks\alpha}(x,\ \theta)=\psi_{ks\alpha}(\theta)\phi_{\alpha}(x,\ \theta)$$

采用非绝热算符方法可导出跃迁矩阵元

$$H'_{fi}=\int\psi^{+}_{k's'\alpha'}(\theta)\left\{-\frac{h^2}{2I}\varphi^{+}_{\alpha'}\left(\frac{\partial^2\varphi_{\alpha}}{\partial\theta^2}+\frac{2\partial\varphi_{\alpha}}{\partial\theta}\frac{\partial}{\partial\theta}\right)\mathrm{d}^3x\right\}\psi_{ks\alpha}(\theta)\mathrm{d}\theta$$

此式不仅适用于纯构象跃迁，也适用于诸如施主-受主跃迁等电子态发生变化的情况，并容易推广到光致构象-电子跃迁和(静电)场致构象-电子跃迁。

用构象动力学理论处理蛋白质折叠，把它看成是由一系列构象跃迁组成的多步的复杂过程，便可建立蛋白质折叠的量子理论[41-42]。假定对折叠的一个基元过程用上面的跃迁矩阵元公式进行计算，对于 N 扭转模跃迁，最后得到速率

$$W=0.37\times10^{-87}\exp\frac{\Delta E}{2k_BT}\left(\sum_j I_j\right)^{-1/2}\left(\sum_j\frac{a_j}{I_j}\right)^{2}\ \mathrm{s}^{-1}$$

式中，$a_j\simeq O\ (1)$；I_j 为第 j 个模原子基团的转动惯量，$\mathrm{g\cdot cm^2}$；ΔE 为各个模的始末态扭转势能差 δE_j 之和。据此，即可得到蛋白质折叠中的成核和坍缩时间；对于 N=4，此时间为 μs 至 0.1ms[43]。速率对时间的 $N^{3/2}$ 依赖表示跃迁的合作性，这可用来解释 α 螺旋和 β 折叠等规则结构在折叠的早期出现。跃迁速率对转动惯量的依赖可用于讨论折叠的接触序。接触序用折叠中残基非局域接触数及其序列距离定义，从经验上已知这是决定折叠速率的关键因子[44-45]，上述理论对此给出了机制的解释。最近本课题组利用速率 W 对转动惯量的依赖，计算出了各个蛋白质的折叠速率，与实验资料很好地一致。笔者还得出了跃迁速率对温度 T 的依赖关系为

$$\mathrm{W}\sim\exp\left(\sum_j\frac{\delta G_j}{2k_BT}\right)T^{-1/2}\qquad\delta G_j=\delta E_j+k_BT\ln\frac{\omega_j}{\omega'_j}$$

其中，ω_j 和 ω' 为第 j 模始末态频率，与化学反应率与温度的 Arrehnius 关系明显不同，此关系已为新近的蛋白质折叠速率实验资料所证实[46]。

以上工作首次发表于 1995 年，通过计算得到了蛋白质折叠基元过程的时标；并导出了折叠速率与温度转动惯量等各种物理量的依赖关系，解释或预见了一些蛋白质折叠的实验。国际上，尽管对蛋白质折叠的研究已经进行了很多，但关注蛋白质折叠速率，则是近几年的事，似乎国外迄今还没有

见到从量子理论出发的工作。

5 细胞开关和熵产生

噬菌体是最简单的生命。本研究组研究了 Lambda 噬菌体操纵基因和调控蛋白相互作用网络及溶原态/裂解态转变特性的动力学，通过调控蛋白和操纵基因 40 个结合态及转录翻译过程动力学分析，得到一组微分方程，进行线性稳定性分析，讨论各类奇点的特性，研究噬菌体感染大肠杆菌的溶原态/裂解态转变，证明这个转变必须通过两次分岔——从单稳态到含 2 稳定焦点 1 鞍点的三点态的分岔以及从三点态到另一单稳态的分岔实现。这种转变是典型的生物学开关，从物理学观点看则是相变。并且从 Poincare-Hopf 拓扑指数定理证了这种相变特性的深层次意义和普遍性[47]。

鉴于噬菌体的溶原态/裂解态转变代表一类重要开关，笔者小组研究了开关中几类奇点的熵产生率，证明溶原态(焦点态)的熵产生率低于裂解态(焦点态)，解释了溶原态比裂解态具有更高的稳定性；证明二者都低于鞍点态和分岔处的多重奇点态。这些分析对研究生物学开关可能具有较普遍意义。

已知平衡邻近的线性区中定态的熵产生率最小。活细胞如同一台化学引擎，循序衔接的酶催化反应实现了高效率的能量和物质的转换。生物化学家把活细胞“系统是按各部分和过程的最经济原则运转的”看作生命机体的分子逻辑。从物理原理的角度看，这个“最经济原则”就是“熵产生最小”。上面对溶原态/裂解态转变中熵产生率的讨论与熵产生最小原理一致，并期望这个原理可以推广到活细胞演化过程中远离平衡的定态[35]。

熵产生的重要性还在于热力学熵与 Shannon 信息量的关系。决定熵的微观态密度遵守 Liouville 方程，而决定信息量的概率分布遵守 Kolmogorov 方程，二者涉及的自由度不同，是投影关系[48]。最近在文献[49]中讨论了信息量的产生和扩散。

笔者小组对比计算了正常细胞和癌细胞的熵产生，证明了通常正常细胞具有较低的熵产生率；这说明了细胞分化过程中熵产生随结构而变化。研究干细胞分化过程中的熵产生率的变化将是一个有趣的问题。从理论上估算了外加电场引起细胞的熵产生变化，一定的电场可以改变正常细胞和癌细胞熵产生率的相对大小，从而改变熵流，再致改变信息流的方向[48,50]。

薛定谔在《什么是生命》一书中提出的“生命以负熵为生”的原理，被认为是生命的热力学基础。但看来这还不完全。如果把两句话合在一起，“生命以负熵为生，生命以熵产生最小而活”，可能就构成了比较完整的生命的热力学图画[15]。

6 结语

基因组信息学是理论生物学的龙头。数千个基因组的测序完成，海量资料数据呼唤着生命科学的理性和实证性的更完美的结合。本文远非完整的评述，仅从作者接触到和工作过的某些侧面，以密码—序列—结构—功能为生命信息运行的主线，讨论了若干基因组信息学相关的基本问题。基因组序列是一部天书，不仅要获取和展现它，更重要的是读懂它；而没有理论基础和理论指导，就会迷茫和丧失远见。大宇宙和小宇宙是统一的，活有机体和无生命自然界都是宇宙的宠儿，它们有统一的规律。我们不期望生命的规律还原为无生命自然界的物理规律，但相信二者是协调的，是相互补充的。理论生物学应以寻找和展开生命信息运行的特有规律为己任。尽管目前它还只是生命科学的一个小分支，但未来有可能为生物科学的理性化和定量化做出更大贡献，正如历史上理论物理学曾经为物理学所做出的贡献那样。

参考文献(References)

[1] Gilbert W. Towards a paradigm shift in biology [J]. *Nature*, 1991, 349 (6305): 99-100.

[2] Pullman B, Pullman A. Quantum biochemistry[M]. New York: Wiley Interesci, 1963.

[3] Murray J D. Mathematical biology[M]. Berlin: Springer, 1989.

[4] Waterman M S. Introduction to computational biology[M]. London: Chapman & Hall, 1995.

[5] Mount D W. Bioinformatics: Sequence and genome analysis[M]. NY: Cold Spring Harbor Laboratory Press, 2001.

[6] Kitano H. Foundations of systems biology [M]. Cambridge, MA: MIT Press, 2001.

[7] Luo L F, Li H. The statistical correlation of nucleotides in protein-coding DNA sequences[J]. *Bulletin of Mathematical Biology*, 1991, 52(3): 345-353.

[8] Luo L F, Lee W J, Jia L J, *et al*. Statistical correlation of nucleotides in a DNA sequence[J]. *Physical Review E*, 1998, 58(1): 861-871.

[9] 罗辽复. 生命进化的物理观[M]. 上海: 上海科技出版社, 2000.
Luo Liaofu. Physical Aspects of Life Evolution [M]. Shanghai: Shanghai Science and Technology Press, 2000.

[10] Grosse I, Herzel H, Buldyrev S V, *et al*. Species independence of mutual information in coding and noncoding regions [J]. *Physical Review E*, 2000, 61(5): 5624-5629.

[11] Yu Z G, Jiang P. Distance, correlation and mutual information among portraits of organisms based on complete genomes[J]. *Physics Letters A*, 2001, 286(1): 34-46.

[12] Otu H H, Sayood K. A divide-and-conquer approach to fragment assembly[J]. *Bioinformatics*, 2003, 19(1): 22-29.

[13] Bauer M, Schuster S M, Sayood K. The average mutual information profile as a genomic signature[J]. *BMC Bioinformatics*, 2008, 9(1): 48.

[14] Luo L F, Gao Y, Lu J. Information-thoretic view of sequence organization in a genome [R/OL]. Quantitative Biology: Genomics, ArXiv: q-bio/1004.3843, 2010, http://arxiv.org/abs/1004.3843.

[15] 罗辽复. 物理学家看生命[M]. 长沙: 湖南教育出版社, 1994.
Luo Liaofu. A physicist looks at the life [M]. Changsha: Hunan Education Publisher, 1994.

[16] Reinert G, Schbath S, Waterman M S. Probabilistic and statistical properties of words: An overview[J]. *Journal of Computational Biology*, 2000, 7(1-2): 1-46.

[17] Hao B, Qi J. Prokaryote phylogeny without sequence alignment: From

avoidance signature to composition distance[J]. *Journal of Bioinformatics and Computational Biology*, 2004, 2(1): 1–19.

[18] Sims G E, Jun S R, Wu G A, *et al*. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions[J]. *PNAS*, 2009, 106(8): 2677–2682.

[19] Lu J, Luo L F, Zhang Y. Distance conservation of transcription regulatory motifs in human promoters [J]. *Computational Biology and Chemistry*, 2008, 32(6): 433–437.

[20] Luo L F, Jia M W. Messenger RNA information: Its implication in protein structure determination and others[M]//J Feng, J Jost, M Qian. Networks: From Biology to Theory. London: Springer, 2007: 291–308.

[21] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project[J]. *Nature*, 2007, 447: 799–816.

[22] Gregory T R, Nicol J A, Tamm H, *et al*. Eukaryotic genome size databases [J]. *Nucleic Acids Research*, 2007, doi:10.1093/nar/gkl828.

[23] **罗辽复. 垃圾** DNA **和信息生物学**[J]. **科学**, 2006, 58: 24–28. Lou Liaofu. *Science*, 2006, 58: 24–28.

[24] Luo L F. Law of genome evolution direction: Coding information quantity grows[J]. *Frontiers of Physics in China*, 2009, 4(2): 241–251.

[25] Kauffman S. Investigation[M]. London: Oxford University Press, 2000.

[26] Luo L F, Bai G Y. Maximum information principle and evolution of nucleotide sequences [J]. *Journal of Theoretical Biology*, 1995, 174: 131–136.

[27] Jin H Y, Luo L F, Zhang L R. Using estimative reaction free energy to predict splice sites and their flanking competitors [J]. *Gene*, 2008, 424 (1–2): 115–120.

[28] Zhang L R, Luo L F. Splice site prediction with quadratic discriminant analysis using diversity measure [J]. *Nucleic Acids Research*, 2003, 31 (21): 6214–6220.

[29] Chen W, Luo L F, Zhang L R. The organization of nucleosomes around splice sites[J]. *Nucleic Acids Research*, 2010, 38(9): 2788–2798.

[30] Lu J, Luo L F, Zhang L R, *et al*. Increment of diversity with quadratic discriminant analysis: An efficient tool for sequence pattern recognition in bioinformatics[J]. *Open Access Bioinformatics*, 2010, in press.

[31] Luo L F. The degeneracy rule of genetic code [J]. *Origins of Life*, 1988, 18(1–2): 65–70.

[32] Luo L F. The distribution of amino acids in the genetic code[J]. *Origins of Life*, 1989, 19(6): 621–631.

[33] Luo L F, Li X Q. Coding rules for amino acids in the genetic code – The genetic code is a minimal code of mutational deterioration [J]. *Origins of Life*, 2002, 32(1): 23–33.

[34] Luo L F, Li X Q. Construction of genetic code from evolutionary stability [J]. *Biosystems*, 2002, 65(2–3): 83–97.

[35] Luo L F. Theoretic-physical approach to molecular biology [M]. Shanghai: Shanghai Science and Technology Publisher, 2004.

[36] Luo L F. A unified theory on construction and evolution of the genetic code[R/OL]. Quantitative Biology: Other Quantitative Biology, ArXiv: q-bio/0908.3067, 2009, http://arxiv.org/abs/0908.3067.

[37] Freeland S J, Wu T, Keulmann N. The case for an error minimizing standard genetic code[J]. *Origins of Life*, 2003, 33(4–5): 457–477.

[38] Knight R D, Freeland S J, Landweber L F. Rewiring the keyboard: Evolution of the genetic code[J]. *Nature Reviews Genetics*, 2001, 2: 49–58.

[39] Santos M A S, Moura G, Massey S E, *et al*. Driving change: The evolution of alternative genetic codes[J]. *Trends in Genetics*, 2004, 20: 95–101.

[40] Luo L F. Conformational dynamics of macromolecules [J]. *International Journal of Quantum Chemistry*, 1987, 32(4): 435–450.

[41] Luo L F. Conformation transitional rate in protein folding[J]. *International Journal of Quantum Chemistry*, 1995, 54(4): 243–247.

[42] Luo L F. Protein folding as a quantum transition between conformational states[J/OL]. Quantitative Biology: Biomolecules, ArXiv: q-bio/0906.2452, 2009, http://arxiv.org/abs/0906.2452.

[43] Qiu L, Pabit S A, Roitberg A E, *et al*. Smaller and faster: The 20 residue Trp-cage protein folds in 4 μs [J]. *Journal of the American Chemical Society*, 2002, 124(44): 12952–12953.

[44] Plaxco K W, Simons K T, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins [J]. *Journal of Molecular Biology*, 1998, 277(4): 985–994.

[45] Ivankov D N, Finkelstein A V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure[J]. *PNAS*, 2004, 101(24): 8942–8944.

[46] Yang W Y, Gruebele M. Rate-temperature relationships in λ-repressor fragment λ_{6-85} folding[J]. *Biochemistry*, 2004, 43(41): 13018–13025.

[47] Ding H, Luo L F. Kinetic model of the lysogeny/lysis switch of phage λ [J]. *Chinese Physics Letters*, 2009, 26(9): 098701.

[48] Luo L F. Entropy production in a cell and reversal of entropy flow as an anticancer therapy[J]. *Frontiers of Physics in China*, 2009, 4(1): 122–136.

[49] **邢修三. 动态统计信息理论**[J]. **中国科学** G **辑**, 2005, 35(4): 337–368. Xing Xiusan. *Science in China, Series G*, 2005, 35(4): 337–368.

[50] Luo L F, Molnar J, Ding H, *et al*. Ultrasonic absorption and entropy production in biological tissue[J]. *Diagnostic Pathology*, 2006, 1: 35.

(责任编辑 朱宇)

·学术动态·

“第七届中国植物病害化学防治学术研讨会”征稿

中国植物病理学会主办的“第七届中国植物病害化学防治学术研讨会” 将于 2010 **年** 11 **月** 20 **日在海口市召开。会议征集反映我国植物病害化学防治领域的最新成果,以研究论文为主。**

征文截止时间:2010 **年** 9 **月** 20 **日。**

联系方式:江苏省南京市卫岗南京农业大学植物保护学院(210095),**陈长;电话:**025-84395641;**传真:**025-84395641;**电子信箱:**ccj100cn@yahoo.com.cn。

会议网址: http://www.cspp.org.cn/newsdata/20105219400.html。