数值模拟蛋白质的三级结构及 Levinthal 悖论

钟伟（武汉大学 武汉，湖北 430072）
学号:2015301020164

摘要：蛋白质三级结构主要是靠氨基酸侧链之间的疏水相互作用，氢键，范德华力和静电作用维持的，在二级结构的基础上借助各种次级键卷曲折叠成特定的球状分子结构的构象。本文利用 python 建立一个简化二维热力学模型来模拟一条肽链的折叠，该模型下纯热力学折叠过程概率可利用正则系综计算。文中对折叠过程的长度、能量、进行了计算，并讨论了氨基酸分子间相互作用强度及水温对该过程影响。最后从动力学情况简单介绍了最新的蛋白质折叠研究方法。

关键词：蛋白质三级结构、蛋白质折叠、热力学过程、分子间作用力、蒙特卡洛法、python、Levinthal 悖论、

引言

蛋白质分子是经过氨基酸分子间脱水缩合反应形成的高分子化合物，氨基酸分子间羟基和羧基脱水缩合形成肽链，肽链盘曲折叠形成蛋白质[1].蛋白质具有折叠性可以折叠成不同形状。肽键的作用力比氨基酸侧链之间的疏水相互作用、氢键、范德华力和静电作用比要大的多，且是方向可完全变化的分子键[2][3].有利我们建立第一部分的二维链条式模型。第二部分我们通过程序的运行计算了长度及能量在一个蒙特卡洛模拟中的变化，并讨论规律，之后我们试图调整参数运行，讨论折叠能量及长度随温度的变化，并类比 Ising 铁磁性物质二级相变并讨论，最后让参数不合常理观察现象并讨论。第三部分我们简单讨论了 Levinthal 悖论，看看再什么情况下能让折叠的 Levinthal 时间接近实际的生物学过程所花费的时间，最后简单介绍了 2017 年 science 刊登的关于计算机与实验循环反馈解码折叠的新方法。

Ⅰ. 科学背景及模型的建立

本部分主要讨论蛋白质三级结构，蛋白质的三级结构决定了蛋白质分子的生物性能，因为其决定了其他分子能否和其它物质结合及其可以进入什么样的大分子中[2]。例如酶的配适性，见 Figure1.

而蛋白质的一级结构就是蛋白质多肽链中氨基酸残基的排列顺序，也是蛋白质最基本的结构。它是由基因上遗传密码的排列顺序所决定的。各种氨基酸按遗传密码的顺序，通过肽键连接起来，成为多肽链，故肽键是蛋白质结构中的主键，肽键的强度远大于范德华力等，我们因此认为在三级折叠中不考虑其成键能量，因此把一个氨基酸分子间无范德华力氢键的形态设为能量 0(如直链)，氨基酸分子间的作用力将由 Ji，j 表征第 i 个和 j 个氨基酸分子间的作用能。则某个位形的总能量为：

$$\mathbf{E} = \sum \boldsymbol{\delta m}, \boldsymbol{nJ}_{\boldsymbol{A(m),A(n)}} \qquad (1)$$

表示对氨基酸分子间能量两两配对求和，δmn 为强度因子，我们将对其形式在接下来的模型中做简化。

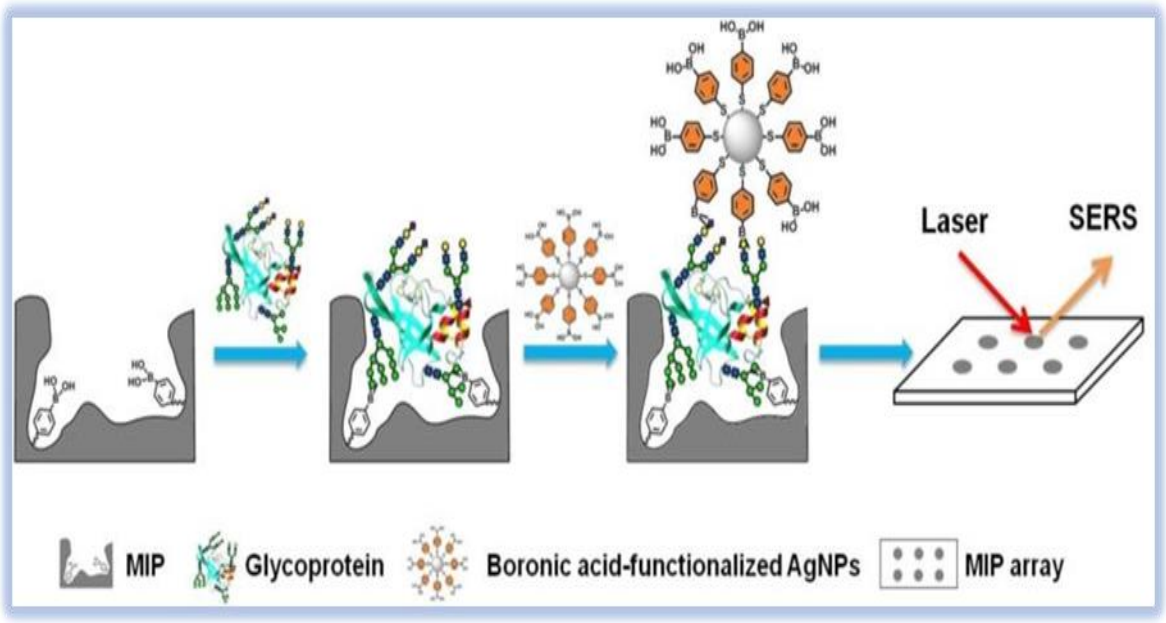

Figure1 蛋白质分子间的配适性，特定空间进入特定分子

将熵引入生命体

生物熵是表征生命活动过程的一种度量。根据热力学第二定律，自然界演化是从有序到无序，而生命的发生、演化、生长过程显然是从无序到有序，生命现象是高度有序的，生命使其内部的熵降低。一个开放系统的熵不一定增加，它可以从外界引入“负熵”。生命正是一个开放系统，由于引入负熵流，总熵变也可以减少，形成有序结构[4]。如一个开放系统，绿色植物通过太阳供给的能量，将无序的无机物（大气中的二氧化碳和土壤中的水等）转化为有序的营养物质，实现局部的熵的减小[5]。但若只研究多肽链形成二级结构的过程， 20 世纪 60 年代，Anfinsen 基于还原变性的牛胰岛 RNase 蛋白在不需其他任何物质帮助下，仅通过去除变性剂和还原剂就使其恢复天然结构的实验结果，提出了“多肽链的氨基酸序列包含了形成其热力学上稳定的天然构象所必需的全部信息”的“自组装学说”[4]。 Anfinsen 的“自组装热力学假说”得到了许多体外实验的证明，尤其是一些小分子量的蛋白。但是另一方面，体内蛋白质的折叠往往需要有其他辅助因子的参与，并伴随有 ATP 的水解供能，这表明蛋白质的折叠不仅仅是一个热力学的过程，显然也受到动力学的控制，而且能量输入暗示了部分蛋白质的组装过程可能也是远离平衡态的熵减过程。

而本文作为一个简单的讨论，考虑结构简单的多肽链，认为仅仅依靠热力学行为，不考虑生命体的动力学体系，即可达到熵最大，最稳态[7]。即认为给定一个多肽链的氨基酸序列，经过时间演化，其能量最低态就是在生物体中发挥功能的蛋白质构型。

二维格点阵模型及实现方法

模型中，不允许两个氨基酸占据同一个格点，从而多肽链不能交叉。由于分子间作用力较弱，仅当两个在序列上不相邻的氨基酸占据相邻两个格点时才会计算其分子间相互作用，则上文（1）式的$J_{i,j}$表示相互作用能前方的强度因子δ只有当序列上不相邻的氨基酸占据相邻两个格点时为 1，

否则为 0. 可以看出一个节点最多只能走斜　　角的四个方位。

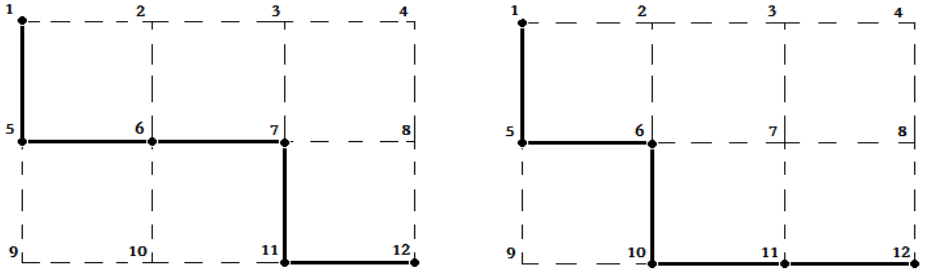

Figure 2 氨基酸空间位置改变的示意图。方形格子的顶点代表氨基酸可能占据的位置，黑色实线段代表某一时刻的蛋白质构型，圆点代表蛋白质的单体氨基酸，黑色实线段代表相邻氨基酸之间的肽键

自由能 F 定为 0，则在正则系综中由配分函数 Z=$e^{-Ei/kT}$ [7]，对一个状态 i，其概率为

$$\mathbf{Pi} = \frac{e^{-Ei/kt}}{\sum_i e^{-Ei/kt}} \qquad (2)$$

关于蛋白质折叠变化的统计判定，我们定义跃迁概率 Pt(i→ j),表示从 Ei 跃迁到 Ej 的概率。则能量为 Ei 的状态跃迁到 Ej 的状态总概率为 Pt(i→ j)Pi。根据刘维尔定律，在一个系综中观察这个稳定的体系，每个态的概率应该不随时间而演变，系综分布函数ρ只是 E 的函数 [7]。

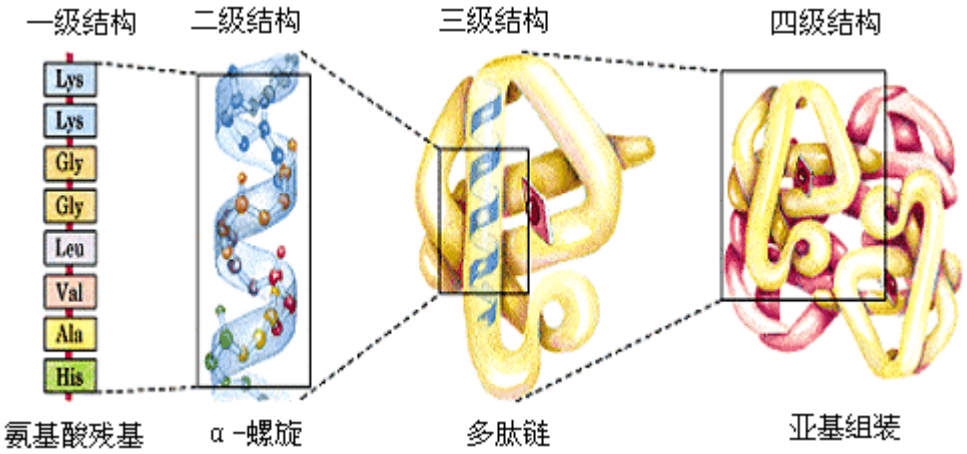

Figure 3 蛋白质的一级到四级结构示意图 三级结构决定了基本的生物功能诺贝尔奖得主 Anfinsen 认为每一种蛋白质分子都有自己特有的氨基酸的组成和排列 顺序，由这种氨基酸排列顺序决定它的特定的空间结构。具有完整一级结构的多肽或蛋白质，只有当其折叠形成正确的三维空间结构才可能具有正常的生物学功能. 如果这些生物大分 子的折叠在体内发生了故障，形成错误的空间结构，不但将丧失其生物学功能，甚至会引 起疾病.蛋白质异常的三维空间结构可以引发疾病，疯牛病、老年性痴呆症、囊性纤维病变、 家族性高胆固醇症、家族性淀粉样蛋白症、某些肿瘤、白内障等等都是“折叠病”。[8]，是生命科学领域的前沿课题之一。不仅具有重大的科学意义，而且在 医学和在生物工程领域具有极大的应用价值。

对我们建立的模型进行规定，每次跃迁只能发生在相邻能级间，从而有等式关系

$$\mathbf{P}(\mathbf{i} \to j)\boldsymbol{Pi} = \boldsymbol{P}(j \to i)\boldsymbol{Pj} \qquad (3)$$

可以得到 Pt(i→ j) = $e^{-(Ej-Ei)/kT}$，且 Pt<1,则有(4)式：

$$p_t = \begin{cases} 1 & \Delta E < 0 \\ e^{-\Delta E/kT} & \Delta E < 0 \end{cases} \quad (4)$$

我们的模型就依赖于 (4) 式的变化，通过蒙特卡洛法产生一系列随机数并进行能量的判断看是否跃迁。我们定义蛋白质分子首尾两端的距离记为 L，用来衡量蛋白质分子链的折叠程度。至此物理模型及相关量已经定义，可以开始整理算法了。

掩码的结构

初始化 J【20，20】矩阵，矩阵元素值随机在一定范围
建立一个折叠过程类 FLODING.CLASS
  初始化蛋白质链（一级结构）NP.ARRAY[20]数组，在 1~20 个整数间随机产生。
  初始化位形 POSITIONS（）产生一条直连
  定义方法 ENERGE(输入当前位形) 利用（1）式计算体系
  定义方法 L（）计算首尾两端的长度
  定义方法 MONTEKLOR()
    定义一个计步变量 COUNTER 为蒙特卡洛步
      当 COUNTER<蒙特卡洛步长
      生产一个 1~20 随机数 N 选定蛋白质链具体节点
      产生一个走位坐标（共四种走位方式）随机值
      计算出位形 NEWPOSITION
        若 N=1 或 20 即在首尾两端
          若 NEWPOSITION 与原 POSITION 有节点重合 COUNTER 加一
          否则
            若新位形节点的产生是否会使肽键断裂（相邻节点距离大于一断裂，只用判断一个节点 COUNTER 加一
            否则
              若新旧能量差<0, 记录新位形，L，能量，及所在步骤
              否则 产生新 0~1 随机浮点数
                若能量差大于随机数 COUNTER 加一
                否则记录新位形，L，能量，及所在步骤，COUNTER 加一
        若 N 取其他数
          若 NEWPOSITION 与原 POSITION 有节点重合 COUNTER 加一
          否则
            若新位形节点的产生是否会使肽键断裂（相邻节点距离大于一断裂，要判断左右两个节点） COUNTER 加一
            否则
              若新旧能量差<0, 记录新位形，L，能量，及所在步骤
              否则
                产生新 0~1 随机浮点数
                若能量差大于随机数 COUNTER 加一
                否则记录新位形，L，能量，及所在步骤，COUNTER 加一

II．数值模拟结果

．首先把温度值设为 10 度，J 中值设在-4KT~-2KT。这里 J 值保证氨基酸分子间的相互作用与温度导致的热运动在同一量级。为简单起见并达到测试程序的目的，我们先看 5 个氨基酸分子构成的蛋白质，准确的说只是一个简单肽链，最简单的蛋白质类型是由 30~50 个氨基酸分子构成的，比如 villin headpiece 和 WW domain 经常被拿来用于蛋白质的折叠研究。[9i]

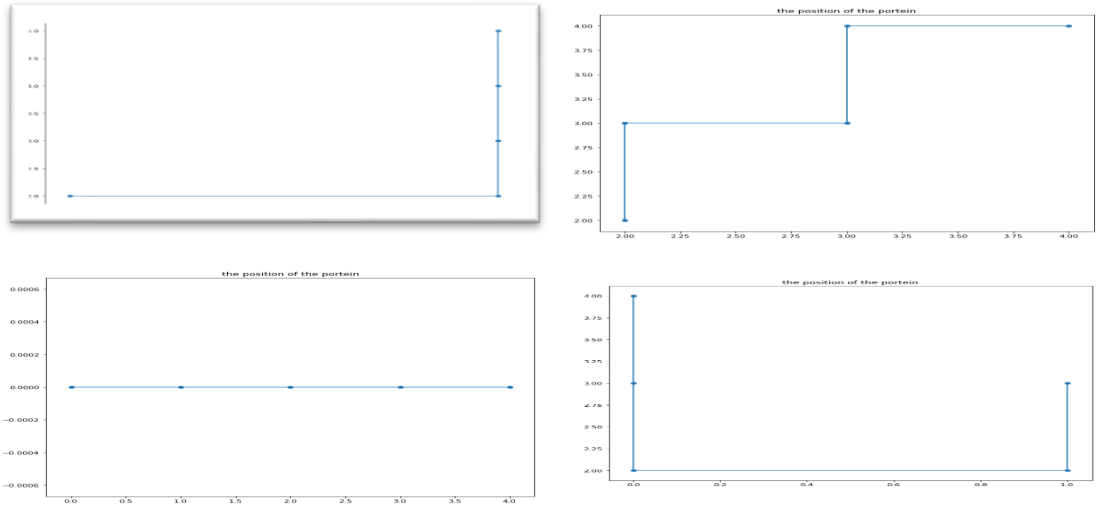

Figure 4 5 个氨基酸分子蒙特卡洛模拟的位形结果。每个点代表一个氨基酸分子，线代表肽键。左下角的图片为初始时的直链，其余的为模拟 2500、5000、7500 步的位形（从左上方依次到右下方）

位形的变化由图像直观地反应出来，最初的直链分子在分子间作用力的情况下趋于能量最低，但显而易见熵不是最大，分子折叠越厉害，越有序，这就是相互作用能克服熵最大得到所谓的负熵的例子[6]

能量的变化与首尾的长度（用来衡量折叠度）变化我们在时间域内画出。

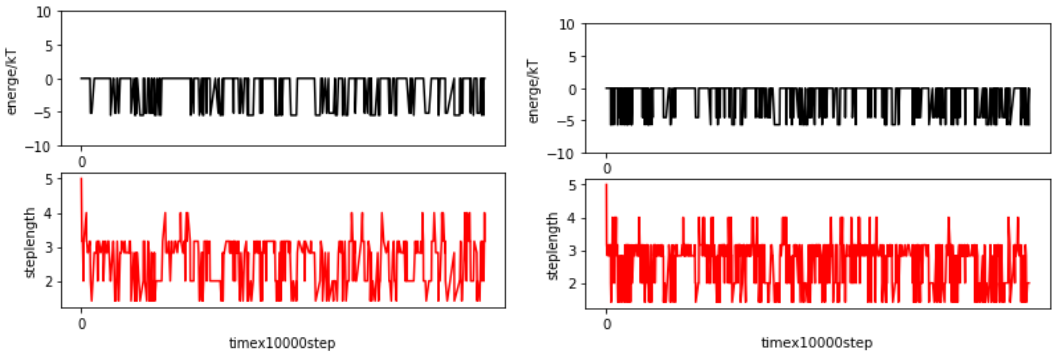

Figure5 蛋白质总能量与长度随时间（蒙特卡洛步数）变化图。黑色折线代表能量，下方红色代表长度。左图为模拟 2500 步结果，右图为模拟 5000 步的结果

能量值稳定在-5~0KT，处于波动中，极差很小，属于正常涨落，平均能量值为负 2.46KT。长度可以看到从起始的 5 快速下降到 2~3 之间，并且稳定波动，平均长度

2.32，与蛋白质快速自折叠吻合。

当尝试其他温度如 15kT 与 50KT，趋势与 10KT 无异（见 Figure6），实际情况下蛋白质在不同温度折叠速度、活性（对应三级结构构型）会有变化，并且超出一定温度范围伴随着失活和解折叠[9]，这是由于温度变化改变氨基酸分子热运动，使范德华力等作用的影响占比变化，结构改变。

如果想解决此问题，直链的 E0 值不能单纯看成 0 常值，E0 应该为温度的函数，并且应采用分段函数的形式，即在一定温度下保持一个常值，超过这个域进入下个域跃变到另一个常值，因为蛋白质有一定温度范围稳定的性质。

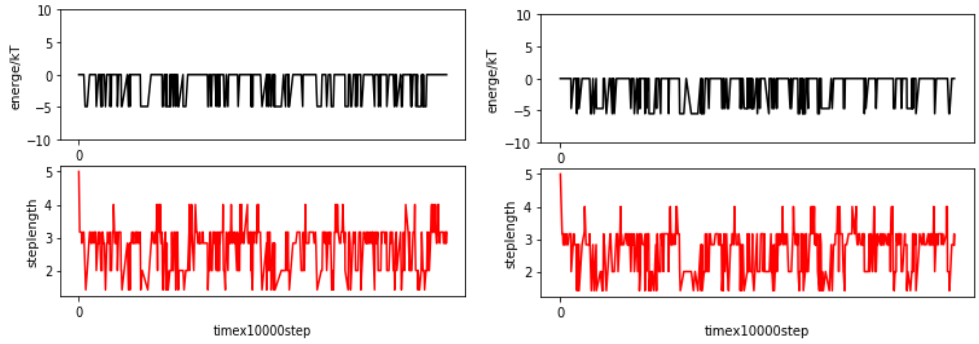

Figure6 左图为 T=1KT,右图为 1000KT 下的模拟，其余条件与 Figure5 同。程序没考虑单个氨基酸的热运动能力导致与实验结果违背。

这个测试说明程序的 T 应该设置的足够小，使分子热运动可以忽略。所以温度范围都在 10KT 以内。

最简单的蛋白质，如有名的 ATP 酶，作为一个被精解的模型荣获 1997 年诺贝尔化学奖被称为动力马达[10]。结构如 figure7 所示

ATP 合酶主要由 F1(伸在膜外的水溶性部分) 和 F0 (嵌人膜内) 组成 。不同物种来源的 ATP 合酶结构不尽相同。[10]。对大肠杆菌 ATP 合酶的研究较为详尽,它的 Fl 含 5 种亚基α_2 、β、γ、δ

ε分子量分别是 55.3、50.3、31.6 、14.9.

.

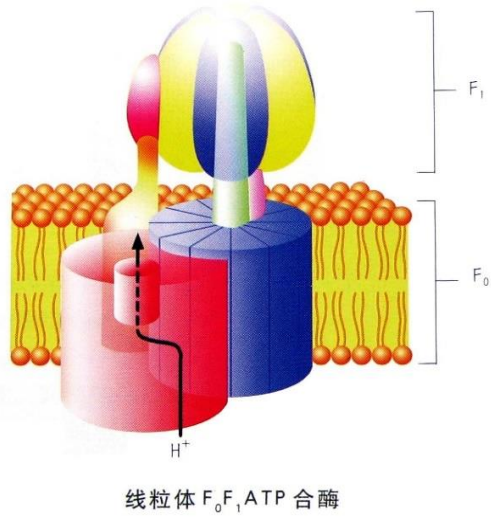

Figure7 线粒体中 ATP 合酶结构

而取 n=30，T=10KT 位形结果如 Figure8

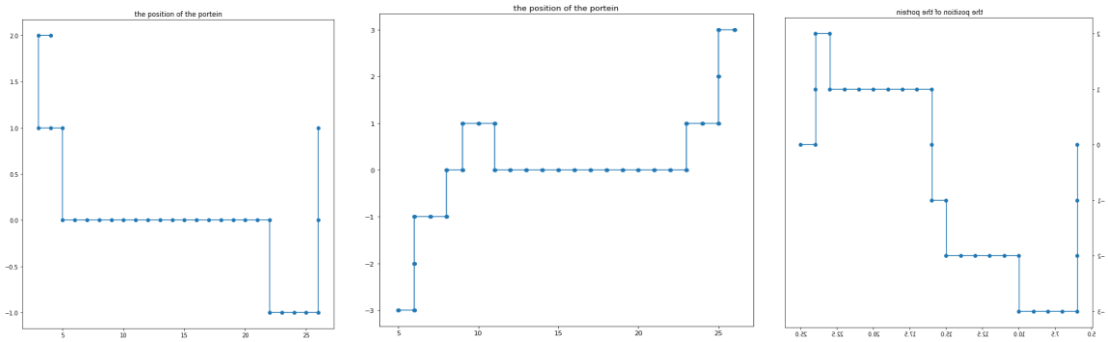

Figure8 三十个氨基酸分子的位形模拟。从左至右分别为 2500、5000、10000 步的位形**可以看出两端的折叠比中间容易实现。**

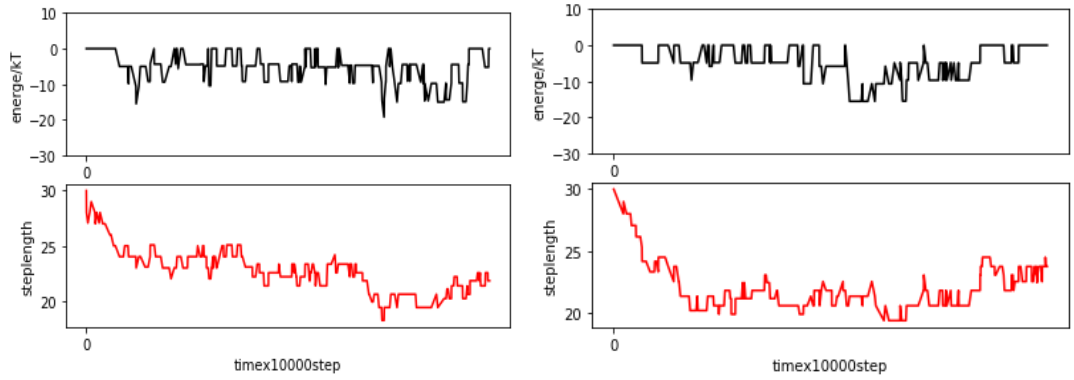

Figure9 30 个氨基酸分子长度及能量随时间变化。左图在 10kT 下右图在 1KT 下

仔细观察稳定后的蛋白质分子长度，1KT 下的蛋白质分子平均长度更小了，即压缩折叠的更厉害了，分子热运动越不剧烈越有利于蛋白质折叠

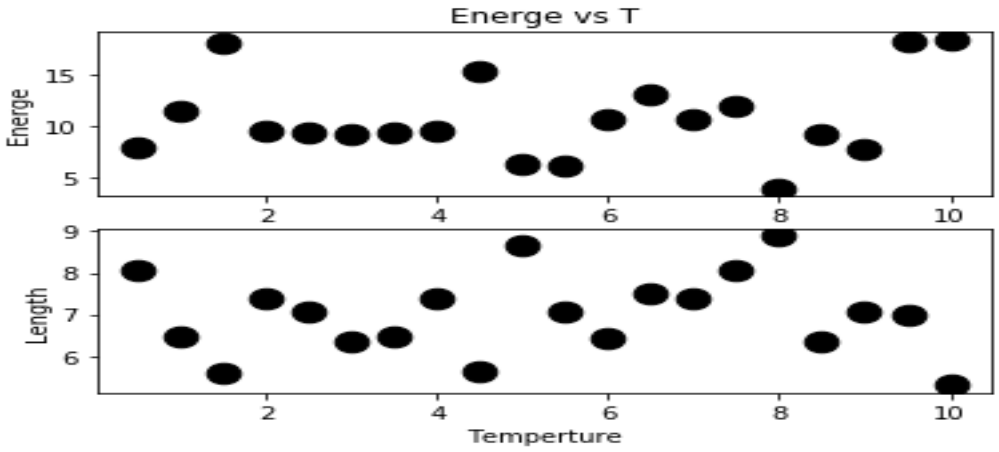

Figure10 15 氨基酸链平均长度与能量随时间变化 每次步长取 5000

进一步观察 0~10KT 范围内的长度能量平均值的变化。如图 10，可惜的的是并未发现如 Ising 模型铁磁性物质转变趋势。点的分布杂乱无章，所以我加大了步数取 500000 不，运行时间更长，很长一时间后结果并没有太大变化

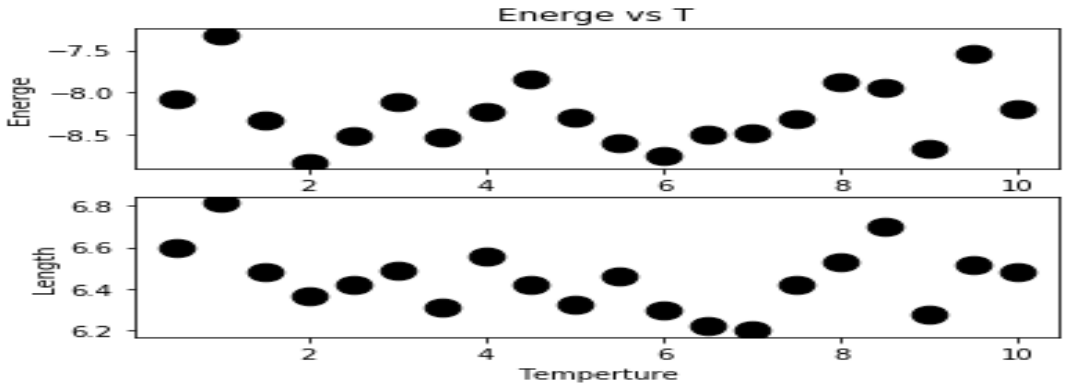

Figure10 步长取 5000 步

具体原因我探究了很久，无果。并且由于点数分布不规则，我没有计算拟合。（代码附在邮件后面，望老师指正，不胜感激）

Ⅲ．Levinthal 悖论及展望

1.在 Levinthal 悖论的标准说明中，连接氨基酸的每个键可以具有几个（例如三个）可能的状态，从而例如 101 个氨基酸的蛋白质可以以 310 = 5×10^{7} 构型存在。即使蛋白质能够以每秒 10^{13} 个，或每年 3×10^{20} 个的速率采样新的配置，也需要 10^{27} 年的时间才能全部尝试。事实上蛋白质折叠能在几秒或更少的时间尺度完成。[3] 所以 Levinthal 认为，随机搜索不是寻找折叠蛋白质正确状态的有效方法。这是悖论。解决矛盾的线索是由道金斯[12]在积累小变化的讨论中提出的。即一种有偏见的搜寻，大大提高了搜素效率。蛋白质分子的这种能量偏好信息存储在蛋白质的序列中，在蛋白质折叠模拟中，势能函数为蒙特卡罗方法[11]和分子动力学方法[12]提供了必要的偏差。

这种方式有点像我们串的模式匹配算法，最开始有 Brute-Force 算法，但效率低，后来提出的 KMP 及其改进算法就是一种带偏好的算法，厉害的是若串越长重复字符越多这种 KMP 算法的效率越高[13]。由于氨基酸种类只有 20 种，则蛋白质分子链越长重复单元越多，无偏好寻找越快。

2.查阅最新的 science 发现，蛋白质折叠开始使用全局分析使用大规模并行设计，合成和测试，建立起一个数据库，用大数据来预测这种折叠，前期与实验数据不断循环反馈，将预测命中率从 7%提高到 47%，并且这种模式会随数据实验量增长不断提高精度。[9]

参考文献

【1】Wikipedia.org/wiki/Protein

【2】Hisao Nakanishi Nicholas J. Giordano. Computational Physics. Tsinghua University Press, second edition

【3】ROBERT ZWANZIG, ATTILA SZABO, AND BIMAN BAGCH Levinthal's paradox Vol. 89, pp. 20-22, January 1992

【4】C. B. Anfinsen. Principles that govern the folding of protein chains. Science, 181(4096):223–30, 1973

【5】Reif. Fundamentals of Statistical and Thermal Physics. Waveland Pr Inc, 2008.

【6】胡辰正 热力学与统计物理学 科学出版社 p56-59

【7】Weiss, G. H. (1967) Adv. Chem. Phys. 13, 1-18

【8】Christopher M.D. Protein Folding and Disease: a view from the first Horizon Symposium Drug Discovery 2003，2：154

【9】Global analysis of protein folding using massively parallel design, synthesis, and

testing Gabriel J. Rocklin 2017 science
【10】杨福愉 ATP 合酶 : 一个最小的蛋 白质分子转动马达 医学分子生物学杂志 , 2005 , 2 (4) : 243- 249 J Mil MQ Bio
【11】Skolnick, J., Kolinski, A. & Yaris, R. (1988) Proc. Natl. Acad. Sci. USA 85, 5057-5061
【12】Honeycutt, J. D. & Thirumalai, D. (1990) Proc. Natl. Acad. Sci. USA 87, 3526-3529
【13】李春葆 数据结构 清华大学出版社 第五版 p134-137