第21卷 第6期
2004年12月

工 程 数 学 学 报
CHINESE JOURNAL OF ENGINEERING MATHEMATICS

Vol. 21 No. 6
Dec. 2004

**文章编号**:1005-3085(2004)06-0862-09

# 生物序列的语义分析与第二密码规则的探索(续)

沈世镒, 余 涛, 开 波, 阮吉寿

(南开大学数学科学学院与 LPMC，天津 300071 )

**摘 要**: 本文继续讨论蛋白质一级结构序列的语义结构，利用组合分析与图论方法讨论 Swiss - Prot 数据库的组合结构，给出 Swiss - Prot 数据库中蛋白质一级结构序列的关键词与核心词的定义、搜索算法与特性参数。并由此给出蛋白质一级结构序列的核心词词典，并由此讨论数据库的复杂性问题、同源蛋白质的分类、预测与比对等问题。

**关键词**: 生物序列结构的语义分析；第二密码规则；蛋白质一级序列结构数据库的组合图论分析；非线性复杂与核心词词典

**分类号**: **AMS**(2000) 92D20 **中图分类号**: O157.1 **文献标识码**: A

## 1 概论

关于生物序列的语义分析的意义与内容含义在文[1]中已有详细说明，在文[1]中我们用信息、统计的方法对蛋白质一级结构序列的语义结构进行分析计算，得到了它们的若干性质，如局部词特性分析、稳定性理论与游程分析问题等。

在信息、统计的方法的使用中，我们注意到信息、统计的方法的基础是多肽链的频数与频率的计算，但当多肽链向量长度适当增加，多肽链向量 $b^n$ 的组合数 $20^n$ 将迅速增长，如 $20^6 = 6.4 \times 10^7$。这时已超过 Swiss - Prot 数据库（20 版）中氨基酸的总数，因此信息、统计的方法就不再有效。

在本文中，我们将利用组合、图论方法继续对蛋白质一级结构序列的语义结构进行分析计算，给出 Swiss - Prot 数据库中蛋白质一级结构序列的关键词的定义与它们的特征参数。关键词与核心词的概念是一种特殊类型的肽链，它可在蛋白质数据库中存在（也就是可在自然界存在）。而且唯一存在。因此关键词与核心词实际上是一种特殊类型的生物标签。

关于生物标签（Singnayures）的概念是生物语义分析的一个常用方法，除了我们在文[1]中提到的小分子库、构象子理论外，国际上许多研究机构都在建立各自的标签数据库，如 PROSITE. Pratt. EMOTIF 等（见文[ 2∼ 10 ]）。它们的方法也各不相同，如 Pratt. EMOTIF 是直接肽链的结构与功能进行分析的标签数据库，而 PROSITE 是利用 PSI - BLAST（见文[11]）对同源蛋白质进行分析比对所产生的同源蛋白质类标签数据。因此，比较这些不同类型数据库的相互关系也是很有意义的。

组合、图论方法的核心问题是序列数据的向量段，在它的长度达到一定长度后，长序列的数据结构就可能形成一种递推关系，这种特性在移位寄存器序列与密码分析中广泛使用，在理论上把它归结为序列数据的复杂度理论与数据结构的布尔（Boolean）函数或德布鲁恩- 古德（de Bruijn - Good）图理论。因此，移位寄存器序列与密码分析中的许多理论与工具都可借

[1]**收稿日期**: 2003-03-26. **作者简介**: 沈世镒(1939年4月生)，男，硕士，教授，研究方向：信息论，生物信息学.
**基金项目**: 天津市南开大学数学科学学院与 LPMC，本文获天津大学、南开大学联合研究项目，刘徽应用数学研究中心与国家自然科学基金(批准号: 10271061; 90208022)资助.

用，但对生物序列的研究与密码分析研究的目的不同，前者是要寻找构成生物序列的词与语的关连性，而密码分析的目的是构造序列的伪随机性。关于组合图论的一般讨论见文[ 12 ∼ 16 ]等文。利用组合、图论的方法还可讨论数据库的复杂性问题、分类问题、剪切与调控问题。本文仅就如何利用核心词对同源蛋白质的分类与预测问题进行探讨。

## 2　关于组合图论的有关记号

关于蛋白质一级结构序列的数学模型与有关的定义、记号在文[1]中已给出，我们不在详细重复，有关布尔函数与德布鲁恩- 古德图的详细理论见文[ 12 ∼ 17 ] 。

### 2.1　组合空间与数据库

记 $Z_q=\{1,2,\cdots,q\}$ 为一个整数集合，代表生物序列的字母表，在蛋白质一级结构序列数据库中取 $q=20$ 分别代表 20 种常用的氨基酸。

为了方便起见，在本文中我们取 $q=23$，这时取 21 为零元，而 $Z_q$ 是一个有限域，它的加、乘运算是以 23 为模的整数加、乘运算。

记 $Z_q^{(n)}$ 是 $Z_q$ 的 $n$ 维乘积空间，它的元 $b^{(n)}=(b_1,b_2,\cdots,b_n)\in Z_q^{(n)}$ 是 $Z_q$ 上的 $n$ 维（或阶）向量。我们又称 $Z_q^{(n)}$ 是 $Z_q$ 的 $n$ 维（或阶）的组合空间。

如文[1]所记，$\Omega$ 是蛋白质一级结构序列数据库，它由 $M$ 条蛋白质组成，这时

$$\Omega=\{C_s,\ s=1,2,\cdots,M\} \tag{1}$$

其中 $C_s=(c_{s,1},c_{s,2},\cdots,c_{s,n_s})$ 是单个蛋白质的一级结构序列，$n_s$ 是第 $s$ 个蛋白质的序列长度，它的分量 $c_{s,i}\in Z_q$ 是常用的氨基酸。如果 $C_s$ 是一个蛋白质的一级结构序列，我们记

$$c_s^{[i,j]}=(c_{s,i},c_{s,i+1},\cdots,c_{s,j}),\quad 1\le i\le j\le N_s \tag{2}$$

是蛋白质 $C_s$ 的一个肽链段，它的长度为 $n=j-i+1$，这时 $c_s^{[i,j]}$ 是 $n$ 阶组合空间的一个向量。

### 2.2　组合空间上的布尔函数

如果 $Z_q^{(n)}$ 是一个组合空间，如果 $f(b^{(n)})$ 是一个 $Z_q^{(n)}\to Z_q$ 的单值映射，那么我称 $f$ 是一个 $Z_q$ 上的 $n$ 阶 $q$ 元的布尔函数。在数学中，对布尔函数有多种表示方法，如列表、组合、函数与图表示等。它们在文[15]中已有详细叙述，我们简述有关记号。

1　组合表示法

列表表示法是人们所熟悉的。组合表示法可用 $Z_q^{(n)}$ 的一组子集合

$$\mathcal{A}_f=\{A_{f,1},A_{f,2},\cdots,A_{f,q}\} \tag{3}$$

来表示，其中

$$A_{f,j}=\{b^{(n)}\in Z_q^{(n)}:\quad f(b^{(n)})=j,\quad j\in Z_q\} \tag{4}$$

那么称 $\mathcal{A}_f$ 是布尔函数的组合表示。这时 $\mathcal{A}_f$ 是 $Z_q^{(n)}$ 的一个分割。

2　函数表示　因为 $f$ 是 $Z_q^{(n)}\to Z_q$ 的映射，所以 $f$ 是一个以 $Z_q^{(n)}$ 为定义域，在 $Z_q$ 上取值的函数，如果 $Z_q$ 是一个有限域，那么布尔函数可用有限域上的加、乘运算计算，它的一般计算式为：

$$f(b^{(n)})=\sum_{j_1,j_2,\cdots,j_n=0}^{q-1}\left(\alpha_{j_1,j_2,\cdots,j_n}\prod_{i=1}^{n}b_i^{j_1}\right) \tag{5}$$

其中 $\alpha_{j_1,j_2,\cdots,j_n}\in Z_q$，且(5)式中各加、乘、幂运算都是 $Z_q$ 域中的运算。

3 图表示 图的一般记号为 $G=\{A,V\}$，其中 $A$ 是它的点集，$V$ 是 $A$ 的一个点偶集合，在图论中称为弧集。以下记 $A$ 中的点为 $a,b,c$ 等，而 $V$ 中的弧为 $A,B,C$ 或 $(a,b),(a,c),(b,c)$ 等。在图的定义中，分有限图、有向图、无向图、子图、倍图与积图等，在本文中我们只讨论有限、有向图。

**2.3 德布鲁尔- 古德图理论**

一种重要的图是德布鲁尔- 古德图（以下简称 DG 图），我们记为 $G_D=\{A_D,V_D\}$，其中

$$A_D=Z_q^{(n)},\quad V_D=Z_q^{(n+1)}, \tag{6}$$

这时 $V_D$ 中的元可写为

$$b^{(n+1)}=((b_1,b_2,\cdots,b_n),(b_2,b_3,\cdots,b_{n+1})), \tag{7}$$

由此可看作 $A_D$ 中的点偶为 $G_D$ 中的弧。

以下称 DG 图的子图也是 DG 图，重要的 DG 图是布尔图，如果 $f$ 是 $Z_q^{(n)}\to Z_q$ 的布尔函数，那么我们称 $G_f=\{A,V_f\}$ 是由 $f$ 决定的布尔图，其中

$$A_f=Z_q^{(n)},\quad V_f=\{(b^{(n)},f(b^{(n)})),\quad b^{(n)}\in A\} \tag{8}$$

这时 $G_f$ 是 $G_D$ 图的子图。布尔图与布尔函数相互确定，在图论中常用的还有算子表示等，对此我们不再介绍。

**2.4 布尔图的性质**

在图论中，相连的弧构成为路，弧与路都有前端（或入口、首端）与后端（或出口或尾端）的定义。如果 $A=(a,b)\in V$ 是弧，那么称弧 $A$ 是 $a$ 的出弧，是 $b$ 的入弧，又称 $a$ 是 $b$ 的先导，$b$ 是 $a$ 的后继。由布尔图的定义知道，DG 图是布尔图的充要条件是每个点最多有一条出弧。

路的出发点又称起点，路的最后到达点为终点，，起点与终点之间的点为中间点。路中弧的数目为路长，我们常用

$$L=\{b_1\to b_2\to\cdots\to b_{k-1}\to b_k\},\quad b_j\in A \tag{9}$$

来表示路，其中 $a_j\in A$ 为图中的点，$a_{j-1}\to a_j=(a_{j-1},a_j)\in V$ 为弧，这时 $a_1$ 为路 $L$ 的起点，而 $a_k$ 为路的 $L$ 的终点。这时路长记为 $l=l(L)=k-1$。

在路的定义中还有圈、单回路、极大圈（又称 Hamilton 回路）、完备回路（又称 Eula 回路）、环与树的一系列定义，对树还分丛、树、反树、干树上的点又分节点、根点、梢点的区分，它们的定义与性质在[ 13∼16 ] 等文中有详细讨论，本文不再重复。

**2.5 由序列产生的 DG 图**

如果 $C=(c_1,c_2,\cdots,c_N)$ 是 $Z_q$ 的一个长度为 $N$ 的序列，其中 $c_j\in Z_q$，记序列 $C$ 中的子向量为：

$$c_i^{(n)}=(a_i,a_{i+1},\cdots,a_{i+n-1}),\quad 1\le i\le N-n+1 \tag{10}$$

如果定义

$$A_n(C)=\{c_i^{(n)}=(c_i,c_{i+1},\cdots,c_{i+n-1}),\quad i=0,1,\cdots,N-n\} \tag{11}$$

那么我们称 $A_n(C)$ 是由序列 $C$ 决定的 $n$ 维向量族。

**定义**2.2 如果 $C$ 是 $Z_q$ 的序列，对任何正整数 $n$，我们定义

$$G_n(C)=\{A_n(C),A_{n+1}(C)\} \tag{12}$$

是由序列 $C$ 决定的 $n$ 阶 DG 图，其中 $A_{n+1}(C)$ 中的元为：

$$c_i^{(n+1)} = (c_i, c_{n+1}, \cdots, c_{i+n}) = ((c_i, \cdots, c_{i+n-1}), (c_{i+1}, \cdots, c_{i+n}))$$

为 $A_n(C)$ 中的点偶，因此 $G_n(C)$ 是由 $C$ 决定的 DG 图。$G_n(C)$ 图有以下性质：

1)　$C$ 是图 $G_n(C)$ 中的一条路，$c_1^{(n)}$ 与 $c_{N-n}^{(n)}$ 点分别是它的起点与终点。

2)　如果 $A_n(C)$ 中的点都不相同，那么图 $G_n(C)$ 无圈，这时图 $G_n(C)$ 是一个布尔图，除了 $c_{N-n}^{(n)}$ 点外，其他点有且只有一条出弧。我们称该 DG 图为由 $C$ 产生的布尔图。这时记

$$f(c_i, c_{i+1}, \cdots, c_{i+n-1}) = c_{i+n}, \quad i = 1, 2, \cdots, N-n \tag{13}$$

为生成 $C$ 的布尔函数。在一般情形下，方程式(13)的解不是唯一的，我们记 $F^{(n)}(C)$ 是使(13)成立的全体解，我们称之为生成 $C$ 的布尔函数族。

3)　如果 $A_n(C)$ 中有相同的点，那么图 $G_n(C)$ 有圈，如果 $c_i^{(n)} = c_j^{(n)}, i < j$，那么

$$c_i^{(n)} \to c_{i+1}^{(n)} \to \cdot \to c_{j-1}^{(n)} \to c_j^{(n)} = c_i^{(n)} \tag{14}$$

构成一个圈。为使图 $G_n(C)$ 无圈，就需要用复杂性理论给以解决。

## 3　数据库的复杂度与关键词理论

为讨论 $G_n(C)$ 在什么条件下是一个布尔图，在密码学中归结为序列的复杂度问题。

**3.1　序列的复杂度问题**　在密码学的序列分析中，复杂度有三种定义，即线性与非线性复杂度与非奇异性复杂度，这些概念在语义的组合分析中经常引用，我们先给出它的定义。

**定义**3.1　如果 $C$ 是已给定的序列，那么有以下定义。

1)　称 $n = K_N(C)$ 是 $C$ 的非线性复杂度，如果满足条件：N − 1 DG 图 $G_n(C)$，是一个布尔图，而 $G_{n-1}(C)$ 不是布尔图。

2)　称 $n = K_L(C)$ 是 $C$ 的线性复杂度，如果在条件 N − 1 中的生成函数 $f$ 是 $Z_q^{(n)} \to Z_q$ 的线性函数成立。

3)　称 $n = K_S(C)$ 是 $C$ 的非奇异性复杂度，如果 $C$ 的生成函数 $f$ 是 $Z_q^{(n)} \to Z_q$ 的非奇异函数。

关于非奇异函数是一类重要的布尔函，它的定义是：在函数式(14)中，当 $(c_{i+1}, \cdots, c_{i+n-1})$ 固定时，$c_i$ 与 $c_{i+n}$ 1 − 1 对应。

在[ 13 - 16 ] 等文中对这三种复杂度给出了一系列性质与计算公式，如它们满足关系式 $K_N(C) \le K_S(C) \le K_L(C)$ 等。对此我们不作详细讨论。

**3.2　数据库的复杂度问题**　在定义3.1中，关于由序列产生图与复杂度的定义，都可推广到数据库的情形。在(1) 的蛋白质一级结构序列数据库中，由每个蛋白质一级结构序列 $C_s$ 所产生图，由(12)式定义为：

$$D_n(C_s) = \{A_n(C_s), V_n(C_s)\}, \quad s = 1, 2, \cdots, M \tag{15}$$

我们定义 $\mathcal{G}_n(\Omega) = \{\mathcal{A}_n(\Omega), \mathcal{V}_n(\Omega)\}$ 为由数据库 $\Omega$ 所产生的图，其中

$$\mathcal{A}_n(\Omega) = \bigcup_{s=1}^{M} A_n(C_s), \quad \mathcal{V}_n(\Omega) = \bigcup_{s=1}^{M} V_n(C_s)$$

由图 $\mathcal{G}_n(\Omega)$ 可同样定义数据库 $\Omega$ 的线性与非线性复杂度与非奇异性复杂度，它们与定义3.1相同，我们不作重复。

**3.3 复杂度在生物学中的意义** 序列复杂度的本质意义是讨论在同一序列中，在什么条件下，不同向量的片段存在递推关系，这种递推关系如何表达及变化，因此它与生物序列的调控、剪切概念密切相关。

在我们对生物序列的计算中发现，序列复杂度计算对单蛋白质序列较为有效，而对数据库 $\Omega$ 的分析不很有效。

**例**3.1 天花粉蛋白可从我国中草药中提取的一种药物蛋白，早期是一种引产药物（见[17]文）。近年来发现对多种癌症与艾滋病有抑制作用，因此受到重视，在 Swiss - Prot 数据库中，二种同源蛋白的名称为：RISA - CHLPN，和 RISA - CHLTR，它们的一级结构分别为：

MIRFLVFSLLILTLFLTAPAVEGDVSFRLSGATSSSYGVFISNLRKALPYERKLYDIPLLRSTLPGSQRYALIHLTNYAD
ETISVAIDVTNVYVMGYRAGDTSYFFNEASATEAAKYVFKDAKRKVTLPYSGNYERLQIAAGKIRENIPLGLPALDSAIT
TLFYYNANSAASALMVLIQSTSEAARYKFIEQQIGKRVDKTFLPSLAIISLENSWSALSKQIQIASTNNGQFETPVVLIN
AQNQRVTITNVDAGVVTSNIALLNRNNMAAIDDDVPMAQSFGCGSYAI
MIRFLVLSLLILTLFLTTPAVEGDVSFRLSGATSSSYGVFISNLRKALPNERKLYDIPLLRSSLPGSQRYALIHLTNYAD
ETISVAIDVTNVYIMGYRAGDTSYFFNEASATEAAKYVFKDAMRKVTLPYSGNYERLOTAAGKIRENIPLGLPALDSAIT
TLFYYNANSAASALMVLIQSTSEAARYKFIEQQIGKRVDKTFLPSLAIISLENSWSALSKQIQIASTNNGQFESPVVLIN
AQNQRVTITNVDAGVVTSNIALLLNRNNMAAMDDDVPMTQSFGCGSYAI

对这二个序列，我们分别记为 $C$，$D$，它们的非线性、非奇异复杂度分别为

$$K_N(C)=K_N(D)=3,\quad K_N(C,D)=94,\qquad K_S(C)=K_S(D)=4,\quad K_S(C,D)\geq 95 \tag{16}$$

由此可见，数据库 $\Omega$ 的非线性复杂度会变得很大（在Swiss - Prot 数据库中，非线性复杂度可在3000 以上）。这种复杂度加大的原因是在数据库 $\Omega$ 中存在大量同源序列。因此，复杂度分析对同源序列的搜索、变化预测与分析是很有用的。

**3.4 数据库的关键词与核心词分析** 为了用组合方法对蛋白质一级结构数据库作有效的分析，我们提出数据库的关键词与核心词的理论。

1 关键词与核心词的定义

1) 称向量 $b^{(n)}$ 是数据库 $\Omega$ 的 $k$ 重关键词，如果 $b^{(n)}$ 在 $\Omega$ 中出现的频数 $N_\Omega(b^{(n)})=k$。

2) 称 $b^{(n)}$ 是 $\Omega$ 的 $k$ 重核心词，如果 $b^{(n)}$ 是 $\Omega$ 的 $k$ 重关键词，且 $N_\Omega(b^{(n-1)})>k$ 与 $N_\Omega(b_+^{(n-1)})>k$ 同时成立，其中

$$b^{(n-1)}=(b_1,b_2,\cdots,b_{n-1}),\quad b_+^{(n-1)}=(b_2,b_3,\cdots b_n)$$

分别是 $b^{(n)}$ 的前或后 $n-1$ 个分量的子向量。

3) $n$ 阶 1 重关键与核心词，此同时分别简称为关键与核心词。

关键词与核心词是蛋白质一级结构序列的“标签”。这就是，如果 $b^{(n)}$ 是 $\Omega$ 的一个关键词，那么在 $\Omega$ 中有且只有一个蛋白质序列 $C_s$ 包含这个向量。

关键词与核心词又是蛋白质的“分类”。如果 $b^{(n)}$ 是 $\Omega$ 的 $k$ 重关键词，它们所在的蛋白质序列分别为 $C_s$。$i=1,2,\cdots,k$，那么蛋白质 $s_1,s_2,\cdots,s_k$ 具有同一关键词 $b^{(n)}$，我们就可把它门看成是同源（或局部同源）蛋白质。

一个蛋白质 $C_s$ 可能包含多个核心词，因此蛋白质一级结构序列具有多“标签”的特点。

**例**3.2 在例子3.1的天花粉蛋白 RISA - CHLPN 和 RISA - CHLTR 中，长度为6的核心词分别有：

C: RFLVFS, PYERKL, YERKLY, TLPGSQ, TNVYVM, NVYVMG, VYVMGY, YVMGYR, VMGYRA, NGQFET, GQFETP, FETPVV, NMAAID, MAAIDD, AAIDDD, DVPMAQ, VPMAQS, PMAQSF, MAQSFG, AQSFGC

D: IRFLVL, TLFLTT, LPNERK, YIMGYR, VFKDAM, FKDAMR, KDAMRK, NNMAAM, NMAAMD, MAAMDD, VPMTQS, PMTQSF, MTQSFG, TQSFGC

2　关键词与核心词的性质

如果 $b^{(n)}$，$c^{(m)}$ 是二个向量，如果存在 $1 \le i < j \le m$，且 $j-i+1=n$，使 $b^{(n)}=c^{[i,j]}$。那么称向量 $c^{(m)}$ 包含 $b^{(n)}$，且称 $c^{(m)}$ 是 $b^{(n)}$ 的外延，而向量 $b^{(n)}$ 是 $c^{(m)}$ 的内缩。关键词与核心词有以下性质。

1)　如果 $b^{(n)}$ 是在蛋白质序列 $C_s$ 中的一个关键词，那么它在蛋白质 $C_s$ 上的任何外延都是关键词。

2)　在数据库 $\Omega$ 中，所有不同的核心词都互不包含，它们在数据库 $\Omega$ 中出现而且只出现一次。

3)　当核心词的向量内缩时，该核心词就转化为 $k$ 重关键词与 $k$ 重核心词从而形成蛋白质的同源结构树。同源结构树对长度较长的核心词是很有意义的。

3　核心词数据库的搜索计算

对固定的数据库 $\Omega$，我们记它蛋白质 $C_s$ 上的全体核心词为

$$\mathbf{C}_s=\{C_{s,j}\quad j=1,2,\cdots,k_s\} \tag{17}$$

其中 $C_{s,j}$ 是 $C_s$ 中的一个子向量，那么定义

$$\mathcal{C}=\{\mathbf{C}_s,\quad s=1,2,\cdots,M\}=\{C_{s,j}\quad j=1,2,\cdots,k_s,\ s=1,2,\cdots,M\} \tag{18}$$

为数据库 $\Omega$ 中的全体核心词。

由数据库 $\Omega$ 确定的全体核心词 $\mathcal{C}$。可采用以下递推算法完成，该算法与[16]文中的非线性复杂度递推算法相似。

A-1　由数据库 $\mathcal{C}$ 分解成 $\mathcal{C}=\bigcup_{n=1}^{\infty}C^{(n)}$，其中 $C^{(n)}$ 是全体 $n$ 阶核心词，记

$$(C^{(n)})^*=\{b^{(n)}\in Z_q^{(n)}:\quad N(b^{(n)})\le 1\} \tag{19}$$

因为对任何 $b^{(n)}\in C^{(n)}$，总有 $N(b^{(n)})=1$ 成立，所以 $C^{(n)}\subset(\mathcal{H}^{(n)})$。

A-2　取 $n_0$ 为算法的起始长度，使 $(C^{(n_0-1)})^*=\phi$，且 $(C^{(n_0)})^*\neq\phi$，其中 $\phi$ 表示空集。这时对任何正整数 $n$ ,必有

$$[(C^{(n)})^*\otimes Z_q]\bigcup[Z_q\otimes(C^{(n)})^*]\subset(C^{(n+1)})^* \tag{20}$$

成立。我们记 $(C^{(n)})^c=Z_q^{(n)}-(C_{(n+1)})^*$，这时 $(C^{(n)})^c$ 就是使 $N(b^{(n)})>1$ 的全体向量。

A-3　如果 $(C^{(n)})^c$ 意志，那么我们就可计算全体

$$b^{(n+1)}\in[(C^{(n)})^c\otimes Z_q]\bigcup[Z_q\otimes(C^{(n)})^c] \tag{21}$$

元的频数，记 $N(b^{(n+1)})>1$ 的元为 $(\mathcal{H}^{(n+1)})^c$。

A-4　由A-3 的递推计算确定一系列 $(C^{(n)})^c$，$n=n_0,n_0+1,\cdots$，该运算直到 $(C^{(n')})^*$ 为空集为止。

4)　由 $(C^{(n)})^c$ 即可确定 $(C^{(n)})^*$ 与 $C^{(n)}$，最后确定 $C=\bigcup_{n=n_0}^{n'}C^{(n)}$ 就为所求的核心词数据库。

**定理**3.1　对以上核心词数据库的搜索算法有以下性质。

1) 该算法可穷尽 Ω 数据库的全体核心词，也就是 $\mathcal{H} = \mathcal{H}_0$ 成立，其中 $\mathcal{H}_0$ 为由算法A-1 - A-4 所得的数据库。

2) 该算法的计算复杂度为 $O(N)$，其中 $N$ 是数据库 Ω 中氨基酸的总数。

该定理的证明是显然的，我们不作详细论述，利用该算法我们在普通 PC 机上实现了对 Swiss - Prot 数据库的全部核心词的搜索计算。

## 4 组合分析的若干应用问题

利用蛋白质一级结构序列数据库的组合分析可在基因与蛋白质结构分析中得到应用，我们仅就同源蛋白质搜索，比对与剪切预测概述如下。

**4.1 同源蛋白质搜索与比对问题** 在上文中我们已经给出，如果 $b^{(n)}$ 是一个核心词，那么它在数据库 Ω 中出现，而且只出现一次。如果把向量 $b^{(n)}$ 作前后收缩，那么就可在数据库 Ω 中形成各种不同阶数的关键词，这些关键词所在的蛋白质在相当长的片段内具有相同的肽链，由此可以找到这些同源蛋白质的稳定区域。

图4.1 例4.1中核心词与蛋白质结构关系图

**例子**4.1 在 Swiss - Prot（20 版）中，一个长度为 20 的核心词：

GREFSLRRGDRLLLFPFLSPQKDPEIYTE

它所在的蛋白质与起始位置分别为：CAML - MOUSE，373 . 它的长度为7的片段，及这些片段在其他蛋白质中出现的编号与位置分别为：

| 频数 | 片段 | 编号 | 位置 | 编号 | 位置 | 编号 | 位置 | 编号 | 位置 |
|---|---|---|---|---|---|---|---|---|---|
| 4 | RGDRLLL | CAMG-MOUSE | 379 | CAML-HOMAN | 379 | CAML-MOUSE | 380 | CAML-RAT | 380 |
| 4 | GDRLLLF | CAMG-MOUSE | 380 | CAML-HOMAN | 380 | CAML-MOUSE | 381 | CAML-RAT | 381 |
| 4 | LLLFPFL | CAMG-MOUSE | 383 | CAML-HOMAN | 383 | CAML-MOUSE | 384 | CAML-RAT | 384 |
| 4 | LLFPFLS | CAMG-MOUSE | 384 | CAML-HOMAN | 384 | CAML-MOUSE | 385 | CAML-RAT | 385 |
| 4 | LFPFLSP | CAMG-MOUSE | 385 | CAML-HOMAN | 385 | CAML-MOUSE | 386 | CAML-RAT | 386 |
| 3 | FSLRRGD | CAMG-MOUSE | 375 | CAML-MOUSE | 376 | CAML-RAT | 376 | | |
| 3 | PQKDPEI | CAMG-MOUSE | 391 | CAML-MOUSE | 392 | CAML-RAT | 392 | | |
| 3 | SLRRGDR | CAMG-MOUSE | 376 | CAML-MOUSE | 377 | CAML-RAT | 377 | | |

关于 CAML - MOUSE 的同源蛋白质的结构关系如图4.1所示，我们也可对同一蛋白质上多

个核心词进行搜索，则可找到同源蛋白质的多个相同片段。 说明：大写英文字母为氨基酸单字符，数字为氨基酸在蛋白质中的位置点，①，②，③，④为蛋白质的代号，它们分别是：

①：CAMG - HOUSE，②：CAML - HUMAN，③：CAML - HOUSE，④：CAML - RAT

我们由此可以判定，这四种蛋白质可能是同源蛋白质。

**4.2　关于蛋白质序列的剪切问题**　如果我们记例子3.1的蛋白质 RISA - CHLPN 和 RISA - CHLTR 分别为 $C$，$D$，它们的非线性与非奇异复杂度都是4，因此由 $C$，$D$ 所产生的4阶 DG 图 $G_4(C)$，$G_4(C_2)$ 是二支反干树图（见[4] 文定义），这样我们就可讨论它们的剪切问题。

1　比较例子3.1的蛋白质 $C$，$D$。的一级结构二序列全长为 $N=289$，其中有10个位点上的氨基酸不同，而其余位点上的氨基酸相同，不同位点的位置是：7，18，50，63，94，123，139，234，272，280。

2　如图4.2所示，我们把 $C$，$D$ 蛋白质画成二条平行线，在平行线中，相同符号表示相同的氨基酸，而不相同符号表示不同的氨基酸，我们把平行线中不同氨基酸的对应点称为剪切点，不同剪切点将蛋白质切割成若干片段，我们记 $s_i,\ t_i,\ i=1,2,\cdots,k+1$ 分别是二个蛋白质的切割片段，其中 $k$ 是切割点的数目。这样蛋白质 $C$，$D$ 可表示为：

$$C=(s_1\to s_2\to\cdots\to s_k),\quad D=(t_1\to t_2\to\cdots\to t_k) \tag{22}$$

3　因为 $s_i,t_i$ 分别是 $C$，$D$ 中的子向量，它们可分别记为：

$$s_i=(c_{j_i},c_{j_i+1},\cdots,c_{j_{i+1}-1},c_{j_{i+1}}),\ t_i=(d_{j_i},d_{j_i+1},\cdots,d_{j_{i+1}-1},d_{j_{i+1}}) \tag{23}$$

这里 $j_i<j_{i+1}$，且 $c_k=d_k$，$k=j_i+1,j_i+2,\cdots,j_{i+1}-1$，$c_{j_i}\neq d_{j_i}$。这时我们记

$$s_i^*=(c_{j_i},c_{j_i+1},\cdots,c_{j_{i+1}-1},d_{j_{i+1}}),\ s_i'=(d_{j_i},c_{j_i+1},\cdots,c_{j_{i+1}-2},d_{j_{i+1}-1}) \tag{24}$$

我们称 $s_i^*$ 与 $s_i'$ 分别是 $s_i$ 的共轭元与相伴元，对 $t_i$ 同样可定义它的共轭元与相伴元，在组合图论中，可利用共轭元与相伴元进行剪切，这就是用状态 $s_i^*$，来取代状态 $s_i$，或用 $t_i^*$，来取代状态 $t_i$，这种不同组合方式的剪切可能产生不同类型的同源蛋白质。

对此我们有结论如下：如果二个蛋白质有 $k$ 个剪切点，那么每个蛋白质可切成 $k+1$ 段，这样不同的剪切组合有 $2^{k+1}$ 种。这些剪切组合都有可能形成新的同源蛋白质。

由此我们可以预测，同源的天花粉蛋白质可能有 $2^{11}\sim 2000$ 种。这种同源蛋白质是否存在，它们的功能特性如何则需要通过生物实验来证实与说明。生物信息学的意义之一就是为生物学家提供试验的范围与内容方向，从而大大缩小试验的规模，序列的剪切可为此提供工具。

图4.2　由剪切产生的同源蛋白质示意图

说明：符号○表示相同的氨基酸。符号×与●表示不同的氨基酸，可称为剪切点。①，②，③，④，⑤，⑥表示由不同剪切点将蛋白质分割的不同肽链。由虚线可将不同肽链连接成新的蛋白质，它们有：

原来蛋白质：①→②→③，　④→⑤→⑥。由剪切产生新的蛋白质：

①→②→⑥，　④→⑤→③，　①→⑤→⑥，

①→⑤→③，　④→②→③，　④→②→⑥。

因此由2个剪切点可预测产生8个同源蛋白质。

## 参考文献：


[1] 沈世镒. 生物序列的语义分析与第二密码规则[J]. 工程数学学报，2004;21(5):665-674

[2] Hart RK, Royyuru AK, Stolovitzky G, Califano A. Syatematic and Fully Automated Identification of Protein Sequence Patterns[J]. J Comp Biol, 2000;7:3/4585-600

[3] Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W. Lipman D J. Gapped BLAST and PSI - BLAST: a new generation of Protein Database Search Programs[J]. Nuc Acids Res, 1997;25(17)：3389-3402

[4] Bairoch A. PROSITE: A Dictionary of Sites and Patterns in Proteins[J]. Nuc Acids Res, 1991;19:2241-2245

[5] Attwood T L, et al. PRINTS Prepares for the Millennium[J]. Nuc Acids Res 1999;27:220-225

[6] Bateman A, et al. Pfam 3.1:1333 Multiple Alignments and Profile HMMs Match the Majority of Proteins[J]. Nuc Acids Res 1999;27:260-262

[7] Bork P, Koonin E v. Protein Sequence Motifs[J]. Curr Op Struct Biol, 1996;6:366-376

[8] Burkhard Rost. Review: Protein Secondary Structure Prediction Continues to Rise[J]. Journal of Structural Biology, 2001;1-15

[9] Brazma A. Approaches to the Automattic Discovery of Patterns in Biosequence[J]. J Comp Biol, 1998;5:279-305

[10] Califano A. SPLASH: Structural Pattern Localization Algorithm by Sequential Histograms[J]. Bioinformatics, 2000;16:341-357

[11] Schaffer A, Aravind L, Madden T L, Shavarin S, Spouge J L, Wolf Y I, Koonin E V, Altschul S F. Improving the Accuracy of PSI - BLAST Protein Database Searches with Composition - Based Statistics and Other Refinements[J]. Nucleic Acids Research, 2001;29:2994-3005

[12] Golomb S W. Shift Regieter Sequences[M]. Holden - Day Inc San Francisco, 1967

[13] Chan A H, Games R A, Key E L[J]. J Combin Theory Ser A, 1982;33:233-246

[14] 万哲先，代宗铎，刘木兰，冯宁. 非线性移位寄存器[M]. 科学出版社，1978

[15] 沈世镒. 组合密码学[M]. 浙江科技出版社，1992

[16] 汪猷主编. 天花粉蛋白[M]. 科学出版社，2000

[17] 沈世镒，余涛，开波. 蛋白质核心词数据库[R]. http://mathbio.nankai.edu.cn


# Semantics Analysis of Biological Sequences and Explore of Second Cipher Rules


SHEN Shi-yi, YU Tao, KAI Bo, RUAN Ji-shou

(School of Mathematical Science and LPMC, Nankai University, Tianjin 300071 )



**Abstract:** We consider the semantics structure problem of biological sequences, continune, using the combinatorial analysis method. We give the definitions, search algorithm and characteristic parameters of key words and kernel words. We obtain a dictionary of kernel words for primary structure sequence of protein, and consider the nolinear complexity problem of protein and database, classification, prediction and alignment problem of homologous proteins.



**Keywords:** semantics analysis of biological sequences; second cipher rules; combinatorial analysis of primary structure database of proteins; nolinear complexity and dictionary of kernel word