Hans Journal of Computational Biology 计算生物学, 2017, 7(3), 31-38
Published Online September 2017 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2017.73004

Hans汉斯

Clustering of G Protein-Coupled Receptor Sequences Based on Factor Model

Hua Wang*, Fenglan Bai, Liwei Liu

College of Sciences, Dalian Jiaotong University, Dalian Liaoning
Email: *1123943421@qq.com

Received: Oct. 9th, 2017; accepted: Oct. 20th, 2017; published: Oct. 24th, 2017

Abstract

G protein-coupled receptors (GPCRs) is a family of peptide proteins, and it is of great theoretical and practical value to clustering analysis of GPCRs. In this paper, the eigenvector representations of protein sequences are given by the classification and physicochemical properties of amino acids. On the basis of this, dimensions of characteristic vectors of the protein sequences are reduced by factor analysis and obtain factor model. The factor model is used to analyze the similarity of 40 G protein-coupled receptor sequences, simultaneously carrying out the clustering analysis. Better results provide a new approach for analyzing and comparing GPCRs.

Keywords

Protein Sequence, Eigenvalue, Factor Model, Cluster Analysis

基于因子模型对G蛋白偶联受体序列进行聚类

王　华*，白凤兰，刘立伟

大连交通大学理学院，辽宁　大连
Email: *1123943421@qq.com

收稿日期：2017年10月9日；录用日期：2017年10月20日；发布日期：2017年10月24日

摘　要

G蛋白偶联受体(G Protein-Coupled Receptors, GPCRs)是一肽类膜蛋白家族，对GPCRs序列进行聚类分析有着重要的理论意义和应用价值。本文根据氨基酸的分类及其物化性质给出了蛋白质序列的特征向量表示，在此基础上用因子分析法对蛋白质序列的特征向量进行降维得到了因子模型，进而利用因子模型

*通讯作者。

文章引用: 王华, 白凤兰, 刘立伟. 基于因子模型对 G 蛋白偶联受体序列进行聚类[J]. 计算生物学, 2017, 7(3): 31-38.
DOI: 10.12677/hjcb.2017.73004

分析了40个GPCRs序列的相似性，并进行聚类分析，得到了较好的结果，为分析比较GPCRs序列提供新的手段。

关键词

蛋白质序列，特征向量，因子模型，聚类分析

1. 引言

每个细胞信号的传递都是借助细胞膜的不同种类的受体，将细胞外的信号传递到细胞内。G 蛋白偶联受体(G Protein-Coupled Receptors, GPCRs)就是一个因能结合和调节 G 蛋白活性而得名的超级膜蛋白家族。作用于 GPCRs 的信号物质，通过影响细胞的 GPCRs 而对 G 蛋白质起作用。因此 GPCRs 被认为是相似的分子机制而起作用。GPCRs 在信号传导中的重要作用，不仅有助于了解细胞信号的传导机制、阐明疾病的致病机理，而且对药物的研究提供新的思路，GPCRs 的功能失调会引发许多疾病，如疼痛、色盲症、哮喘等。通过调节有关 GPCRs 介导的信号传导，可以治疗高血压、紧张和消化道溃疡等病症。大部分药物可通过靶向作用于 GPCRs 而达到治疗的效果，所以 GPCRs 在制药领域成为重要的药物作用靶标。根据 GPCRs 的序列差异，准确地聚类 GPCRs 序列有着很重要的理论意义和应用价值[1] [2]。

蛋白质序列相似性分析是蛋白质序列聚类的关键所在。经典的相似性算法有 Needleman-Wunsch 算法、Smith-Waterman 算法、接触度量矩阵法、矩阵法、T-coffee 算法、SIM 算法、基于氨基酸物化性的拓扑指数方法和基于 LZ 复杂度等方法[3] [4] [5] [6]。尽管这些算法都有它们各自的优点，但是计算量都比较大。在序列比对算法中，仅用相同或不同来说明两个残基之间的关系，无法描述残基替换对结构和功能的影响程度；矩阵法是构造一个适当的数值矩阵来描述蛋白质序列，然后选取矩阵的不变量把蛋白质序列之间的比较转化成矩阵不变量之间的比较，并且把初始的蛋白质字符串序列转换为特征向量，这些特征向量的维数可以按照自己的愿望进行选择。通过对这些不变量的比较来确定蛋白质序列的相似与不相似[7]-[12]。虽然这个方法简洁快速，但是在蛋白质序列转化成特征向量的过程中，总会丢失了一些生物信息；同样在 LZ 复杂度法中也只考虑了结构信息而忽略了氨基酸的物化性质对蛋白质的结构和功能的影响[13]，为此促使人们试图寻找其它更有效的方法来比较蛋白质序列的相似性。

本文中，首先在氨基酸的物化性质表征蛋白质序列的基础上，把蛋白质序列转化成 11 维的特征向量；其次，根据 20 种氨基酸的极性、非极性、疏水性、亲水性将其分为四类：极性且亲水性(pq)、极性且疏水性(pr)、非极性且亲水性(sq)和非极性且疏水性(sr)，将这四类氨基酸两两连接得到 16 个特征子列，并计算了 16 个特征子列在蛋白质序列中出现的频率，利用此频率将蛋白质序列转化成 16 维特征向量；最后，用因子分析法把蛋白质序列的特征向量进行降维得到因子模型，进而利用因子模型分析 40 个 G 蛋白偶联受体序列的相似性，并对其进行聚类分析。

2. 因子模型

$X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$是可观测的随机向量，$E(X)=\mu$，$Var(X)=\sum$，$F=(F_1,F_2,\cdots,F_m)^{\mathrm{T}}(m<p)$是

不可观测的随机向量， $E(F)=0$， $Var(F)=I_m$，又设 $\varepsilon=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p)^{\mathrm{T}}$ 与 F 不相关，且 $E(\varepsilon)=0$，$D(\varepsilon)=\mathrm{diag}(d_1^2,d_2^2,\cdots,d_p^2)=D$。设 X 满足：

$$\begin{cases} X_1=a_{11}F_1+a_{12}F_2+\cdots+a_{1m}F_m+\varepsilon_1 \\ X_2=a_{21}F_1+a_{22}F_2+\cdots+a_{2m}F_m+\varepsilon_2 \\ \vdots \\ X_p=a_{p1}F_1+a_{p2}F_2+\cdots+a_{pm}F_m+\varepsilon_p \end{cases} \tag{2.1}$$

矩阵方程 $X_{p\times1}=A_{p\times m}F_{m\times1}+\varepsilon_{p\times1}$ 称正交因子模型，其中 $F_1,F_2,\cdots,F_m$ 称 X 的公共因子，$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p$ 称 X 的特殊因子。设 $F_1,F_2,\cdots,F_m$ 分别是均值为 0，方差为 1 的随机变量，即 $D(F)=I_m$；特殊因子 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p$ 分别是均值为 0，方差为 $d_1^2,d_2^2,\cdots,d_p^2$ 的随机变量，即 $D(\varepsilon)=\mathrm{diag}(d_1^2,d_2^2,\cdots,d_p^2)=D$；各特殊因子之间及特殊因子与公共因子之间都是相互独立的，即 $Cov(\varepsilon_i,\varepsilon_j)=0,i\neq j$ 及 $Cov(\varepsilon,F)=0$。a_{ij} 是第 j 个变量在第 i 个公共因子上的负荷，$A=(a_{ij})_{p\times m}$ 是待估系数矩阵(因子载荷矩阵) [14]。

因子分析的目标是找出公共因素及特有的因素，即公共因子与特殊因子。在公因子分析中，特殊因子起到残差的作用，但被定义为彼此不相关且和公因子也不相关。而且每个公因子假定至少对两个变量有贡献，否则它将是一个特殊因子。在开始提取公因子时，为了简便，还假定公因子彼此不相关且具有单位方差。在这种情况下，向量 X 的协方差矩阵 $\sum$ 可以表示为：

$$\sum=D(X)=D(AF+\varepsilon)=AA'+D \tag{2.2}$$

这里 $D=\mathrm{diag}(d_1^2,d_2^2,\cdots,d_p^2)$， diag 表示对角矩阵。

如果已知 X 协方差矩阵 $\sum$ 和 D，可以很容易地求出 A。根据式(2.2)有：

记 $\Sigma^*=\Sigma-D$，则 Σ^* 是非负定矩阵。若记矩阵 Σ^* 的 p 个特征值 $\lambda_1,\lambda_2,\cdots,\lambda_m,\cdots,\lambda_p$ 且 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m>\lambda_{m+1}=\cdots=\lambda_p=0$，及 $e_1,e_2,\cdots,e_m$ 分别表示 m 个非零特征值所对应的标准化的特征向量，则 Σ^* 的谱分解式为：

$$\Sigma^*=\lambda_1e_1e_1'+\lambda_2e_2e_2'+\cdots+\lambda_me_me_m'=\left(\sqrt{\lambda_1}e_1,\sqrt{\lambda_2}e_2,\cdots,\sqrt{\lambda_m}e_m\right)\left(\sqrt{\lambda_1}e_1,\sqrt{\lambda_2}e_2,\cdots,\sqrt{\lambda_m}e_m\right)' \tag{2.3}$$

只要：

$$A=\left(\sqrt{\lambda_1}e_1,\sqrt{\lambda_2}e_2,\cdots,\sqrt{\lambda_m}e_m\right) \tag{2.4}$$

就可以求出因子载荷矩阵 A。从而求得 $d_i^2=1-\sum_{t=1}^{m}a_{ti}^2,i=1,2,\cdots,p$，这时 A 和 $D=\mathrm{diag}(d_1^2,d_2^2,\cdots,d_p^2)$ 为因子模型的一个解，这个解就称为主因子解。

因子模型被估计后，还必须对得到的公因子 F 给出一种明确的解释，它用来反映在预测每个可观察变量中这个公因子的重要性，这个公因子的重要程度就是在因子模型矩阵中相应于这个因子的系数，显然这个因子的系数绝对值越大越重要，而接近 0 则表示对可观察变量没有什么影响。因子解释是一种主观的方法，有时侯，通过旋转公因子可以减少这种主观性，也就是要使用非奇异的线性变换。

设 p 维可观察变量 X 满足因子模型 $X=AF+\varepsilon$，设 Γ 是任一正交阵，则因子模型可改写为：

$$X=A\Gamma\Gamma'F+\varepsilon=A^*F^*+\varepsilon \tag{2.5}$$

其中，$A^*=A\Gamma$，$F^*=\Gamma'F$。

3. 蛋白质序列的特征向量表示

3.1. 蛋白质序列的 11 维、2 维特征向量表示

根据大量的实验结果，我们提取了对 G 偶联受体序列相似性分析有影响的氨基酸的 11 种物化属性[15]，它们是：α_c (α 螺旋的 C 端动力)，C_α (α 螺旋接触面积)，K^0 (可压缩性)，P_β (β 折叠趋势)，R_a (在溶剂中的收缩率)，ΔASA (溶剂可及表面积)，PI (氨基酸的等电点)，ΔG_{hD} (吉布斯自由能变性蛋白水化的变化)，N_m (平均中程接触)，Mu (折射率)，EL (长距离的非键能)，并对其进行了标准化和平均化[15]，见表 1。

根据 20 种氨基酸的标准化和平均化后的物化属性，对 40 个 G 蛋白受体序(https://www.ncbi.nlm.nih.gov/protein/1LMB_4 上下载)中所含的 20 个氨基酸进行统计并计算它们的算数平均数，由于数据篇幅比较大，表 2 只给出了部分 G 蛋白受体序列的算术平均数。

Table 1. 20 properties of amino acids
表 1. 20 种氨基酸的属性取值表

Residue	Property										
	α_c	C_α	K^0	P_β	R_a	ΔASA	PI	ΔG_{hD}	N_m	Mu	El
Ala	0.58	0.15	0.83	0.35	0.32	0.22	0.404	−0.58	0.83	0.34	0.36
Asp	0.97	0.27	0.24	0.13	0.14	0.21	0	−6.1	0.51	0.28	0.09
Cys	0.21	0.25	0.26	0.62	0.21	0.57	0.288	−1.91	0.59	0.84	0.7
Glu	0.9	0.42	0	0	0.26	0.29	0.36	−7.37	0.81	0.41	0.13
Phe	0.34	0.69	0.13	0.76	0.82	0.84	0.339	−1.35	0.69	0.69	0.79
Gly	0.13	0	0.71	0.29	0.23	0	0.401	−0.82	0.22	0	0.43
His	0.09	0.5	0.34	0.38	0.3	0.52	0.603	−5.57	0.69	0.51	0.45
Ile	0.16	0.54	0.34	0.92	1	0.8	0.407	0.4	0.47	0.45	0.87
Lys	0.11	0.69	0.29	0.28	0	0.35	0.872	−5.97	0.67	0.5	0
Leu	0.11	0.46	0.34	0.7	0.69	0.69	0.402	0.35	0.92	0.44	0.66
Met	0.19	0.62	0.39	0.51	0.58	0.84	0.372	−0.71	1	0.51	0.66
Asn	0.3	0.31	0.41	0.39	0.06	0.24	0.33	−6.63	0.55	0.31	0.15
Pro	1	0.19	1	0.14	0.06	0.24	0.442	0.56	0	0.26	0.3
Gln	0.45	0.48	0.28	0.55	0.15	0.4	0.056	−7.12	0.75	0.41	0.19
Arg	0	0.88	0.74	0.42	0.13	0.58	1	−12.78	0.65	0.63	0.47
Ser	0.23	0.15	0.49	0.29	0.11	0.15	0.364	−6.18	0.26	0.15	0.28
Thr	0.48	0.31	0.38	0.62	0.14	0.27	0.354	−3.66	0.26	0.26	0.42
Val	0.13	0.42	0.43	1	0.91	0.58	0.399	0.18	0.33	0.33	0.81
Trp	0.56	1	0.46	0.75	0.76	1	0.39	−4.71	0.61	1	1
Tyr	0.18	0.69	0.09	0.83	0.21	0.82	0.362	−8.45	0.37	0.74	0.66

Table 2. The arithmetic mean of 40 proteins
表 2. 40 种蛋白质的算数平均值

	α_c	C_α	K^0	P_β	R_a	ΔASA	PI	ΔG_{hD}	N_m	Mu	El
Q8MXU2	0.3368	0.4119	0.4275	0.5117	0.3785	0.4486	0.4089	−3.51	0.537	0.3996	0.4723
Q9V4U4	0.3639	0.3814	0.4581	0.4911	0.3447	0.4094	0.4093	−3.3994	0.5257	0.377	0.4443
Q09630	0.3394	0.4199	0.4177	0.5024	0.3656	0.4487	0.4117	−3.6989	0.5459	0.4085	0.4571
P91685	0.3446	0.4199	0.4177	0.5024	0.3579	0.4302	0.4018	−3.6051	0.5295	0.3915	0.4517
P31421	0.3499	0.3943	0.4603	0.5059	0.3704	0.4319	0.4174	−3.3861	0.5494	0.4002	0.4744

根据得到蛋白质序列对应的11维特征向量，并计算11维特征向量的相关矩阵及其特征值，根据特征值累计贡献率提取相应的主因子，以主因子的方差贡献率作为权重得到因子模型如下：

$x_1=-0.7378f_1+0.2902f_2$，$x_2=0.6938f_1-0.5964f_2$，$x_3=-0.4949f_1+0.8335f_2$，$x_4=0.8966f_1-0.2191f_2$，$x_5=0.9301f_1+0.0991f_2$，$x_6=0.8680f_1-0.3875f_2$，$x_7=0.0469f_1+0.0173f_2$，$x_8=-0.0463f_1+0.1128f_2$，$x_9=0.0220f_1-0.0383f_2$，$x_{10}=0.2801f_1+0.3878f_2$，$x_{11}=0.2922f_1+0.3878f_2$

其中 x_1 表示 α_c，x_2 表示 C_α，x_3 表示 K^0，x_4 表示 P_β，x_5 表示 R_a，x_6 表示 ΔASA，x_7 表示 PI，x_8 表示 ΔG_{hD}，x_9 表示 N_m，x_{10} 表示 Mu，x_{11} 表示 EL，

$$\begin{aligned}f_1&=-0.10958x_1^*-0.01489x_2^*-0.0463x_3^*+0.15327x_4^*+0.31804x_5^*-0.04243x_6^*\\&\quad+0.04687x_7^*-0.04615x_8^*+0.02197x_9^*+0.28006x_{10}^*+0.29216x_{11}^*\\f_2&=-0.04775x_1^*-0.15665x_2^*+0.57288x_3^*+0.03451x_4^*+0.25414x_5^*-0.23404x_6^*\\&\quad+0.01727x_7^*+0.11279x_8^*-0.03826x_9^*-0.00144x_{10}^*+0.3878x_{11}^*\end{aligned}$$

$x_i^*\left(i=1,2,\cdots,11\right)$ 是 $x_i\left(i=1,2,\cdots,11\right)$ 的标准化，利用因子模型可得蛋白质序列对应的2维特征向量 $\left(f_1,f_2\right)$。例如，蛋白质序列Q8MXU2对应的二维特征向量 $\left(f_1,f_2\right)=\left(0.5598,-0.0536\right)$。

3.2. 蛋白质序列的 16 维、6 维特征向量表示

根据 20 种氨基酸的极性、非极性、疏水性、亲水性将其分为四类[16]：极性且亲水性 $pq=\{G\}$、极性且疏水性 $pr=\{A,V,L,I,F,P\}$、非极性且亲水性 $sq=\{S,C,N,E,T,Q,K,R,H\}$ 和非极性且疏水性 $sr=\{W,Y,M\}$，将这四类氨基酸两两连接得到 16 个特征子列：$pqpq$，$pqpr$，$prpq$，$pqsq$，$sqpq$，$pqsr$，$srpq$，$prpr$，$prsq$，$sqpr$，$prsr$，$srpr$，$sqsq$，$sqsr$，$srsq$，$srsr$，并计算这些特征子列在蛋白质序列中出现的频率得到蛋白质序列的 16 维特征向量表示。在此基础上，根据得到蛋白质序列对应的 16 维特征向量，并计算 16 维特征向量的相关矩阵及其特征值，根据特征值累计贡献率提取相应的主因子，以主因子的方差贡献率作为权重得到因子模型。利用因子模型可得蛋白质序列对应的 6 维特征向量 $(f_1,f_2,f_3,f_4,f_5,f_6)$。例如，蛋白质序列 Q8MXU2 对应的二维特征向量 $\left(f_1,f_2,f_3,f_4,f_5,f_6\right)=\left(0.0314,0.0750,-0.0189,-0.0911,0.0319,0.2247\right)$。

4. 蛋白质序列相似性及其聚类分析

从 https://www.ncbi.nlm.nih.gov/protein/1LMB_4 上下载了 40 个 G 蛋白受体序列[17]。下面采用向量端点之间的平方和距离来构造相似性矩阵。由于数据过多，40 个 G 蛋白偶联受体序列对应的相似性矩阵不列在本文里。把相似性矩阵输入到 SAS 软件得到了 40 个 G 蛋白受体序列的聚类图，如图 1~图 4 所示。

观察聚类图 1~图 4 易看出，40 个 G 蛋白偶联受体序列中：P97772 和 Q9UGT0，O00222 和 Q3MIV9，Q93564 和 Q622H2，Q6ZMQ2 和 Q14833，P31421 和 Q14416，Q5RAL3、Q9QYS2 和 P31422，Q863I4、O15303 和 P35349，Q93564 和 Q622H2 最相似。根据文献[13]，如 Q68EF4 和 Q14833，P47743 和 P70579，Q9V4U4 和 Q70GQ8，Q5TZ45 和 P31424 是比较相似的。而通过观察图我们发现，只有 16 维向量和 6 维向量所得到的聚类图满足，11 维向量和 2 维向量的到的聚类图只满足一部分。由此可见，根据氨基酸的极性和亲水性得到的蛋白质相似性比较更加合理。而图 2 和图 4 是利用因子分析法，把 40 个 G 蛋白偶联受体序列对应的 11 维和16维特征向量降维到2维和6维得到的聚类图。在图2中Q75QW7，Q90ZF3，Q4RJZ9，Q68EF4，Q9V4U4，Q8NHA9，Q4R3P0 的位置和图 1 的有差别。但是通过参考文献[13]，我们知道 Q75QW7 应该和 Q4REZ5 相近，Q90ZF3、Q4RJZ9 和 Q4REI8 相近，Q14833 和 Q68EF4 相近，Q8NHA9 和 O15303 相近。在图4中Q8MXU2，Q75QW7，Q5RDQ8，Q62916 的位置和图 3 的有差别。而在参考文献[13]中，Q75QW7 和 P91685 相近，Q5RDQ8

和 Q62916、Q14833 和 Q68EF4 相近。由此可见，基于 2 维和 6 维特征向量聚类比 11 维和 16 维特征向量聚类效果较好。再把图 2 与图 4 与文献[13]比较可知，图 4 与文献[13]的结果更一致，这说明，我们选取的 20 种氨基酸的极性和亲水性对 G 蛋白偶联受体序列相似性比较中有一定的影响，总之，用低维特征向量聚类过程简单，时间复杂度低等优势，而且实例验证，基于 2 维特征向量聚类效果也较好。

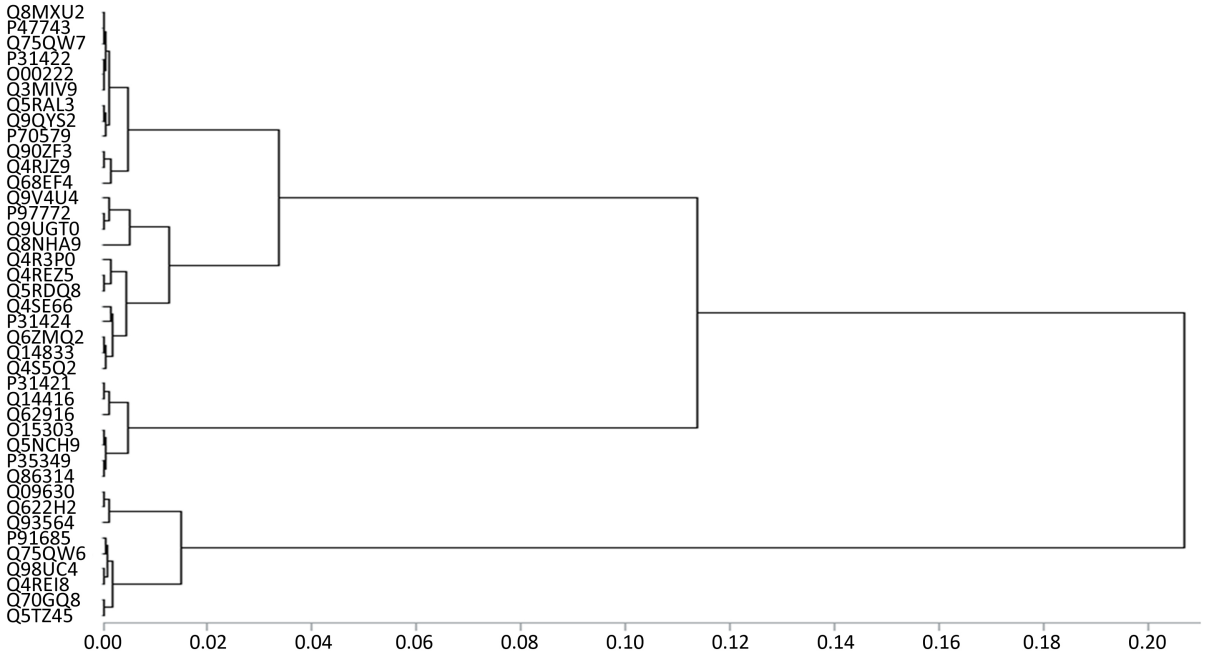

Figure 1. Cluster analysis of 40 G protein-coupled receptor sequences based on 11-dimensional factor vector
图 1. 基于 11 维特征向量的 40 个 G 蛋白偶联受体序列聚类图

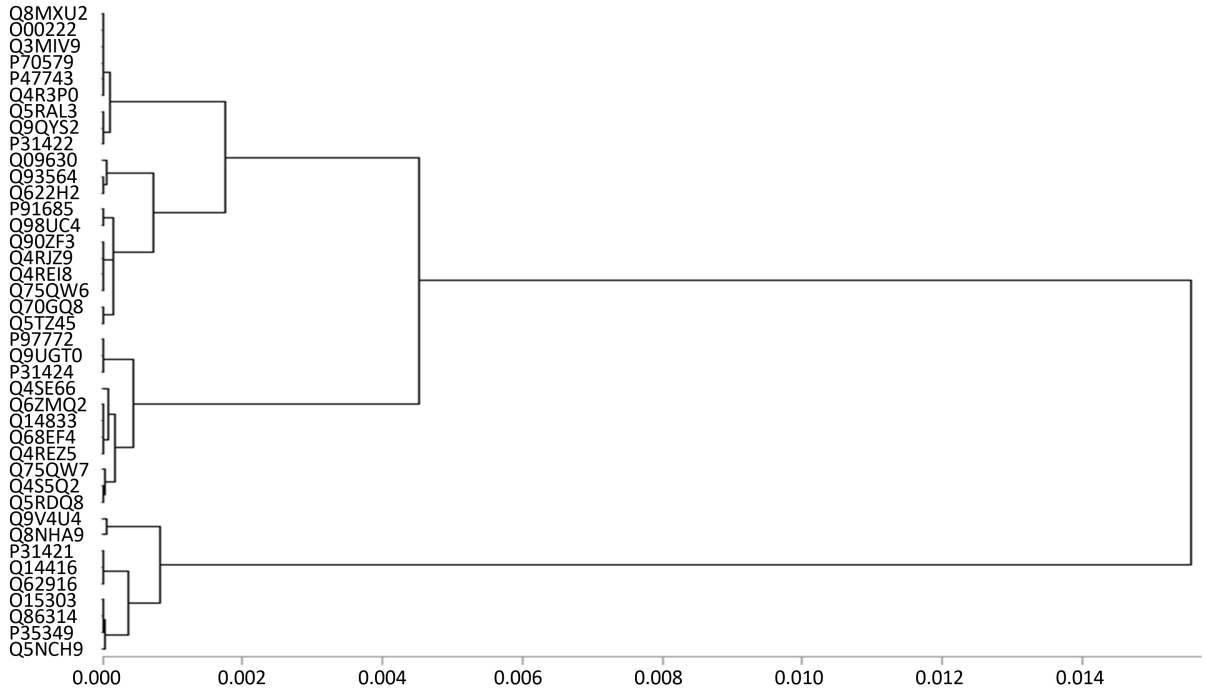

Figure 2. Cluster analysis of 40 G protein-coupled receptor sequences based on 2-dimensional factor vector
图 2. 基于 2 维特征向量的 40 个 G 蛋白偶联受体序列聚类图

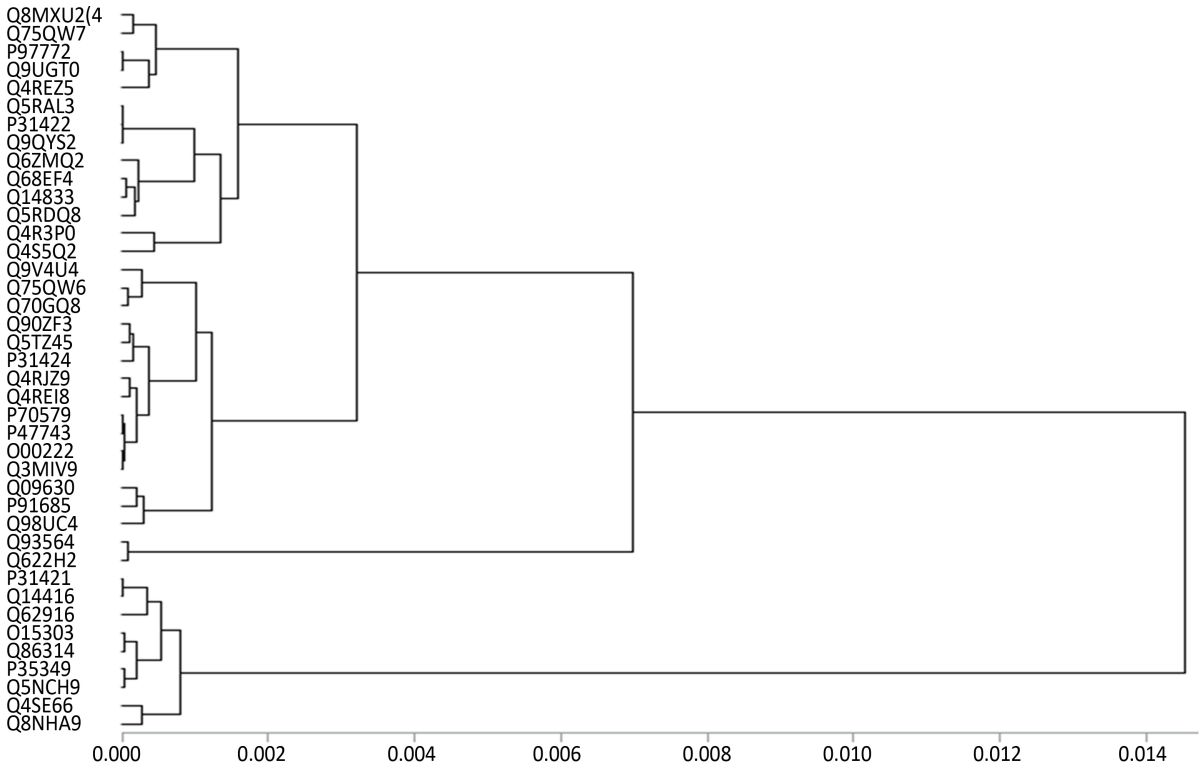

Figure 3. Cluster analysis of 40 G protein-coupled receptor sequences based on 16-dimensional factor vector
图 3. 基于 16 维特征向量的 40 个 G 蛋白偶联受体序列聚类图

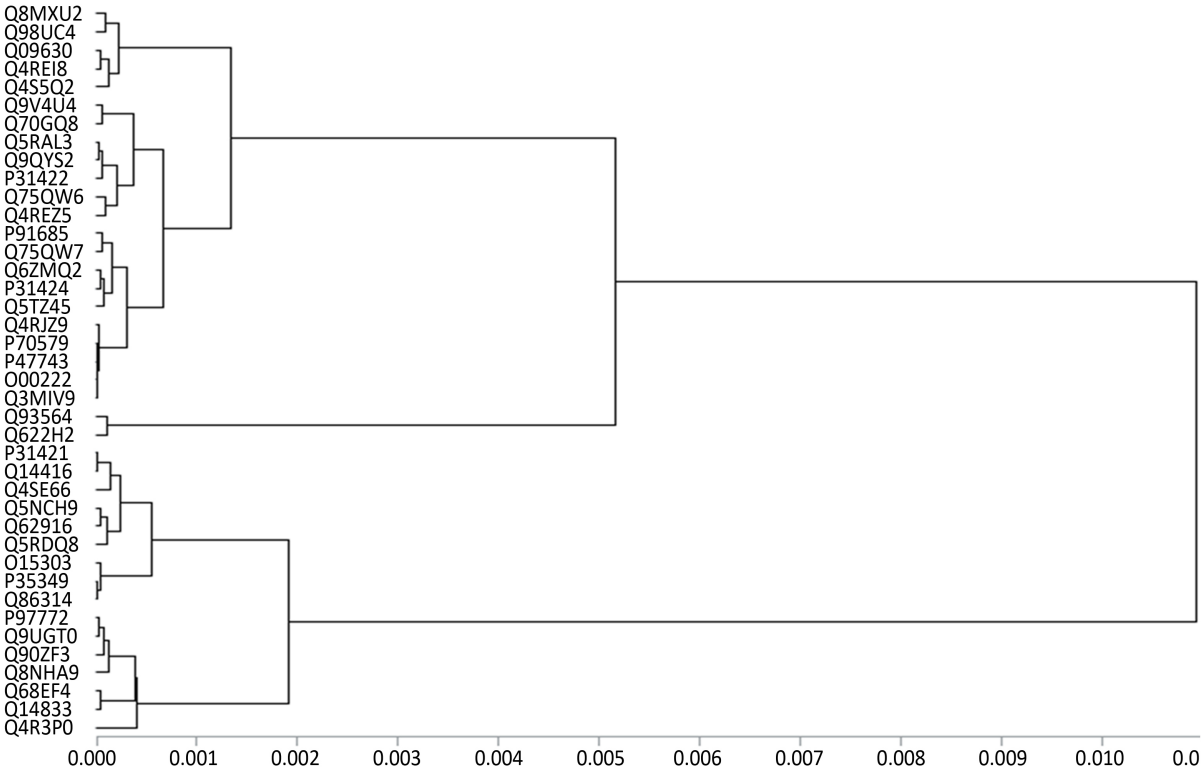

Figure 4. Cluster analysis of 40 G protein-coupled receptor sequences based on 6-dimensional factor vector
图 4. 基于 6 维特征向量的 40 个 G 蛋白偶联受体序列聚类图

5. 总结

因子分析是主成分分析的推广，它是一种降维方法，其目的是用有限个不可观测的隐变量来解释原始变量之间的相关性。在因子模型中，由于因子载荷矩阵不是唯一的，我们利用这一特点可以通过因子的旋转，使得旋转后的因子有更鲜明的实际意义。本文利用因子分析法，对 40 个 G 蛋白偶联受体序列进行相似性分析，并将其聚类，得到的结果验证了本方法的可行性。而且，此方法对生物序列进化的研究具有操作简单、时间复杂度低、对序列的长度没有限制等优点。对于本文中我们选取的氨基酸的 11 种物化性质在不同功能的蛋白质中有没有规律以及分析这些物化性质与蛋白质功能和结构之间的影响都是有待于进一步研究的课题。

基金项目

感谢基金项目：辽宁省教育厅科学研究一般项目(No.L 2015093)对本论文的支持。同时，也要衷心地感谢本文中引用文章的作者。

参考文献 (References)

[1] Bockaert, J. and Pin, J.P. (1999) Molecular Tinkering of G Protein-Coupled Receptors: An Evolutionary Success. *The EMBO Journal*, **18**, 1723-1729. https://doi.org/10.1093/emboj/18.7.1723

[2] Wu, J.S., Ma, X., Zhou, T., *et al*. (2010) Prediction of G-Protein Coupled Receptors and Their Type. *Acta Biochimica et Biophysica Sinica*, **26**, 138-148.

[3] Liu, N. and Wang, T.M. (2006) Protein-Based Phylogenetic Analysis by Using Hydropathy Profile of Amino Acids. *FEBS Letters*, **580**, 5321-5327.

[4] Liu, N. and Wang, T.M. (2006) A Method for Rapid Similarity Analysis of RNA Secondary Structures. *BMC Bioinformatics*, **7,** 493-503. https://doi.org/10.1186/1471-2105-7-493

[5] Lisewski, A.M. and Lichtarge, O. (2006) Rapid Detection of Similarity in Protein Structure and Function through Contact Metric Distances. *Nucleic Acids Research*, **34**, 1-10. https://doi.org/10.1093/nar/gkl788

[6] Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *Journal of Molecular Biology*, **302**, 205-217. https://doi.org/10.1006/jmbi.2000.4042

[7] Stuart, G.W., Moffett, K. and Baker, S. (2002) Integrated Gene and Species Phylogenies from Unaligned Whole Genome Protein Sequences. *Bioinformatics*, **18**, 100-108. https://doi.org/10.1093/bioinformatics/18.1.100

[8] Solovyev, V.V. (1993) Fractal Graphical Representation and Analysis of DNA and Protein Sequences. *Biosystems*, **30**, 137-160. https://doi.org/10.1016/0303-2647(93)90067-M

[9] Das, J., Basu, S., Pan, A. and Dutta, C. (1997) Chaos Game Representation of Proteins. *Journal of Molecular Graphics and Modelling*, **15**, 279-289. https://doi.org/10.1016/S1093-3263(97)00106-X

[10] Randić, M. (2004) 2-D Graphical Representation of Proteins Based on Virtual Genetic Code. *SAR and QSAR in Environmental Research*, **15**, 147-157. https://doi.org/10.1080/10629360410001697744

[11] Balaban, A.T., Randic, M. and Zupan, J. (2004) Unique Graphical Representation of Protein Sequences Based on Nucleotide Triplet Codons. *Chemical Physics Letters*, **397**, 247-252.

[12] Krilov, J. and Randic, M. (1997) Characterization of 3-D Sequences of Proteins. *Chemical Physics Letters*, **272**, 115-119.

[13] Liu, N. and Wang, T. (2007) Comparison of Biological Sequences/Structures and Construction of Phylogenetic Trees. Da Lian University, Da Lian.

[14] 范金城, 梅长林. 数据分析(第二版)[M]. 北京: 科技出版社, 2010: 137-150.

[15] 李欣颖, 白凤兰. 蛋白质序列的混合特征值对折叠速率的影响[J]. 生物信息学, 2014, 12(3): 225-231.

[16] 李巍巍, 李阳, 唐旭情. 不同特征描述下 H1N1 病毒血凝素蛋白质序列的比较分析[J]. 生命科学研究, 2016, 20(2): 119-124.

[17] Bai, F., Gao, H., Liu, L. and Liu, X. (2010) The Similarity Comparison of G-Protein Coupled Receptor Based on Structural Matrix Algorithm. 2010 *International Conference on Computational and Information Sciences*, 653-656, Chengdu, Vol. 12, 17-19.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2164-5426，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：hjcb@hanspub.org