Modeling and Simulation 建模与仿真, 2016, 5(4), 123-130
Published Online November 2016 in Hans. http://www.hanspub.org/journal/mos
http://dx.doi.org/10.12677/mos.2016.54017

Hans汉斯

The Research of Feature Extraction Algorithm by Integrating T-Rank and Softmax Methods

Zhe Liu[1], Chunli Peng[2], Peng Chen[1], Youxi Luo[3*]

[1]School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan Hubei
[2]School of Economic and Management, Hubei University of Technology, Wuhan Hubei
[3]School of Science, Hubei University of Technology, Wuhan Hubei
Email: *youxiluo@163.com

Received: Oct. 19th, 2016; accepted: Nov. 8th, 2016; published: Nov. 11th, 2016

Abstract

The paper proposed a new feature extraction algorithm by integrating T-rank and Softmax for the high dimensional biological data sets, which is more effective than traditional method when dealing with high dimensional data. It can not only extract a very few number of features, but also have fast computing speed. By using of this new algorithm, the paper obtains a high accuracy diagnosis model for psoriasis.

Keywords

High Dimensional, Softmax Algorithm, T-Rank Algorithm, Psoriasis, Gene Expression

一种融合T-Rank和Softmax的特征提取算法研究

刘　哲[1]，彭春力[2]，陈　鹏[1]，罗幼喜[3*]

[1]湖北工业大学电气与电子工程学院，湖北　武汉
[2]湖北工业大学经济与管理学院，湖北　武汉
[3]湖北工业大学理学院，湖北　武汉
Email: *youxiluo@163.com

*通讯作者。

文章引用: 刘哲, 彭春力, 陈鹏, 罗幼喜. 一种融合T-Rank和Softmax的特征提取算法研究[J]. 建模与仿真, 2016, 5(4): 123-130. http://dx.doi.org/10.12677/mos.2016.54017

收稿日期：2016年10月19日；录用日期：2016年11月8日；发布日期：2016年11月11日

摘 要

本文针对高维生物数据特征提出了一种融合T-Rank和Softmax的特征提取算法。该方法比传统特征提取方法在处理高维生物数据更加有效，不仅提取的特征个数较少，而且计算速度快。利用算法本文对高维银屑病基因表达谱数据进行了研究，得到了分类准确率较高的疾病诊断模型。

关键词

高维，Softmax算法，T-Rank算法，银屑病，基因表达谱

1. 引言

近年来，由于科学技术的发展以及基因诊断的进步，人们对高维生物数据有了更深的认识，基因表达谱数据一次性可以获得成千上万个基因片段的表达值，然而很多疾病只与少数几个关键致病基因有关。利用特征选择算法有助于在缺乏先验知识的情况下缩小致病关键基因的候选范围，并深入研究在分子层面上致病机理。目前关于关键特征基因筛选的方法大致可以分为三类：过滤法、缠绕法、混合法[1]。过滤法主要是用指标对基因进行排序筛选，方法简单，但忽略了基因间的相互信息，分类准确性较差。缠绕法主要将特征选择与分类器缠绕在一起，使得选择的特征能有较好的分类准确性，然而该方法对于高维数据计算量极大。混合法则是上述两种的结合。基因表达数据的高维性和冗余性使得基于机器学习的混合法有着较好应用。李霞等[2]较早地提出了一种基于递归分类树的集成特征选择方法 EFST，该方法对不同的分类器都有较好的适应性。李颖新等[3]较早的将支持向量机应用到了肿瘤分类特征基因识别中。吕飒丽等[4]使用决策森林来进行特征选择，再使用人工神经网络作为分类器，获得了很好的分类效果。张飞等[5]在肺鳞状癌细胞发展的特征基因提取中建立了四步筛选方案：相关性筛选、显著性筛选、偏最小二乘算法、基于模式识别分类精度的综合筛选，实证分析显示了多重筛选机制的必要性，构建的分类器对三个集有较好的准确率，重要的是筛选出的特征基因得到了分子生物学层面的解释。银屑病是[6] [7]一种常见的慢性复发性炎症性皮肤病，但是银屑病的病因尚未阐明。本文将针对银屑病基因表达谱数据提出一种新的特征选择算法，并构建银屑病基因诊断的分类模型。

2. 模型与算法

2.1. Softmax 理论模型

我们将已经有 m 个标记了的训练样本作为训练集，特征向量 x 的维度为 $n+1$，即 $x^{(i)} \in R^{n+1}$，最终训练集组成的集合为：$\left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(m)}, y^{(m)}\right)\right\}$。(我们对符号的约定如下：特征向量 x 的维度为 $n+1$，其中 $x_0 = 1$对应截距项)。

在 Softmax regression 中的假设函数如下：

$$h_\theta\left(x^{(i)}\right) = \begin{bmatrix} p\left(y^{(i)} = 1 \middle| x^{(i)}; \theta\right) \\ p\left(y^{(i)} = 2 \middle| x^{(i)}; \theta\right) \\ \vdots \\ p\left(y^{(i)} = k \middle| x^{(i)}; \theta\right) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^{\mathrm{T}} x^{(i)}}} \begin{bmatrix} e^{\theta_1^{\mathrm{T}} x^{(i)}} \\ e^{\theta_2^{\mathrm{T}} x^{(i)}} \\ \vdots \\ e^{\theta_k^{\mathrm{T}} x^{(i)}} \end{bmatrix}$$

其中 θ 矩阵可以写成下面的形式：

$$\theta = \begin{bmatrix} -\theta_1^{\mathrm{T}} - \\ -\theta_2^{\mathrm{T}} - \\ \vdots \\ -\theta_k^{\mathrm{T}} - \end{bmatrix}$$

此时，系统损失函数的方程为：

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)} = j\right\} \log \frac{\mathrm{e}^{\theta_j^{\mathrm{T}} x^{(i)}}}{\sum_{l=1}^{k} \mathrm{e}^{\theta_l^{\mathrm{T}} x^{(i)}}}\right]$$

其中的 1{.}是一个指示性函数，即当大括号中的值为真时，该函数的结果就为 1，否则其结果就为 0。

Softmax regression 中损失函数的偏导函数如下所示：

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[x^{(i)}\left(1\left\{y^{(i)} = j\right\} - p\left(y^{(i)} = j \middle| x^{(i)}; \theta\right)\right)\right]$$

$\nabla_{\theta_j} J(\theta)$ 表示的是损失函数对第 j 个类别的第 l 个参数的偏导。

Softmax regression 中对参数的最优化求解不只一个，每当求得一个优化参数时，如果将这个参数的每一项都减掉同一个数，其得到的损失函数值也是一样的。这说明这个参数不是唯一解，数学公式如下：

$$p\left(y^{(i)} = j \middle| x^{(i)}; \theta\right) = \frac{\mathrm{e}^{(\theta_j - \psi)^{\mathrm{T}} x^{(i)}}}{\sum_{l=1}^{k} \mathrm{e}^{(\theta_l - \psi)^{\mathrm{T}} x^{(i)}}} = \frac{\mathrm{e}^{\theta_j^{\mathrm{T}} x^{(i)}} \mathrm{e}^{(-\psi)^{\mathrm{T}} x^{(i)}}}{\sum_{l=1}^{k} \mathrm{e}^{\theta_l^{\mathrm{T}} x^{(i)}} \mathrm{e}^{(-\psi)^{\mathrm{T}} x^{(i)}}} = \frac{\mathrm{e}^{\theta_j^{\mathrm{T}} x^{(i)}}}{\sum_{l=1}^{k} \mathrm{e}^{\theta_l^{\mathrm{T}} x^{(i)}}}$$

由于在实际的使用过程中一般要加入规则项，加入规则项后的损失函数表达式如下：

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)} = j\right\} \log \frac{\mathrm{e}^{\theta_j^{\mathrm{T}} x^{(i)}}}{\sum_{l=1}^{k} \mathrm{e}^{\theta_l^{\mathrm{T}} x^{(i)}}}\right] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2$$

偏导函数表达式如下所示：

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[x^{(i)}\left(1\left\{y^{(i)} = j\right\} - p\left(y^{(i)} = j \middle| x^{(i)}; \theta\right)\right)\right] + \lambda\theta_j$$

最终通过程序求得的 θ，该矩阵大小为 $k*(n+1)$，k 为分类的类别数，此处 $k = 2$ (针对二分类的问题)，对于输入的数据为 data (即为输入矩阵)，标签 labels 为 $\{1, 2, \cdots, k\}$ 。

2.2. 基于 T 检验理论模型

T-test 检验方法是比较独立样本的一种假设检验方法，此方法的零假设是 H_0 是两总体的均值相等，备择假设 H_1 是均值不等，通过 T 检验可以比较两个总体间的均值是否有着显著区别。

$$t_{ij} = \frac{\overline{x}_i - \overline{x}_j}{\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}},\ i \neq j,\ i \text{ 与 } j = 1, 2, \cdots, r$$

同时有：$P_{ij} = P\left\{t(n-r) > \left|t_{ij}\right|\right\}$，当 $\left|t_{ij}\right| > t_{\frac{\alpha}{2}}(n-r)$ 或 $P_{ij} < \frac{\alpha}{2}$ 时，可以判断 u_i 与 u_j 差异显著。T-rank 即根据所得到的 p-value 对基因特征显著程度进行排序的一种算法。

2.3. 融合 T-rank 的 Softmax 的特征提取算法

本文提出的融合 T-rank 和 Softmax 算法流程图如图 1。

3. 实验材料

3.1. 实验数据及预处理

本实验的两个基因表达谱数据集 GSE14905 [7]，GSE13355 [8] [9]均来自 GEO 数据库(Gene Expression Omnibus) [10]。两个数据总共有 176 个样本，其中 91 个来自银屑病病例样本，85 个来自健康对照组样本。原始.CEL 文件分别经过背景校正，log2 转换和 RMA (Robust Multichip Analysis)算法归一化处理[11]。其中归一化算法采用 Affymetrix Expression Console TM 软件处理(http://www.affymetrix.com/estore/index.jsp)。由于两个数据集来自不同的实验室，我们还利用 DWD(Distance Weighted Discrimination) [12]算法进行了系统偏差的消除。

我们删除了一些总体方差过小判别能力较低的探针集，其中在调用 IQR (Inter-Quartile Range)函数时 var.cutoff 参数设置为 0.5，其它参数为默认参数。所有的过滤步骤应用于 DWD 算法校正之后的数据集，最终有 27336 个探针集的 176 个样本被保留进入下一步的分析。

3.2. 特征选择方法

首先我们考虑进行数据的降维，采用 Deep learning 中的 Softmax 回归算法，该算法是 Logistics 回归算法推广，在求解 Cost function 的时候，能很好的利用 L-BFGS 求得系统的参数，同时，L-BFGS 算法运行速度快。之所以没有直接采用 Logistic 回归是因为在实际应用中发现如果类别是互斥的，那么多采用 Logistics 回归，而类别如果是一样的，则多采用 Softmax 回归算法，本题属于二分类问题，明显类别是不互斥的，综上，我们采用 Softmax 算法进行特征提取，为了保证算法的可靠性，又融合了 T-rank 算法。

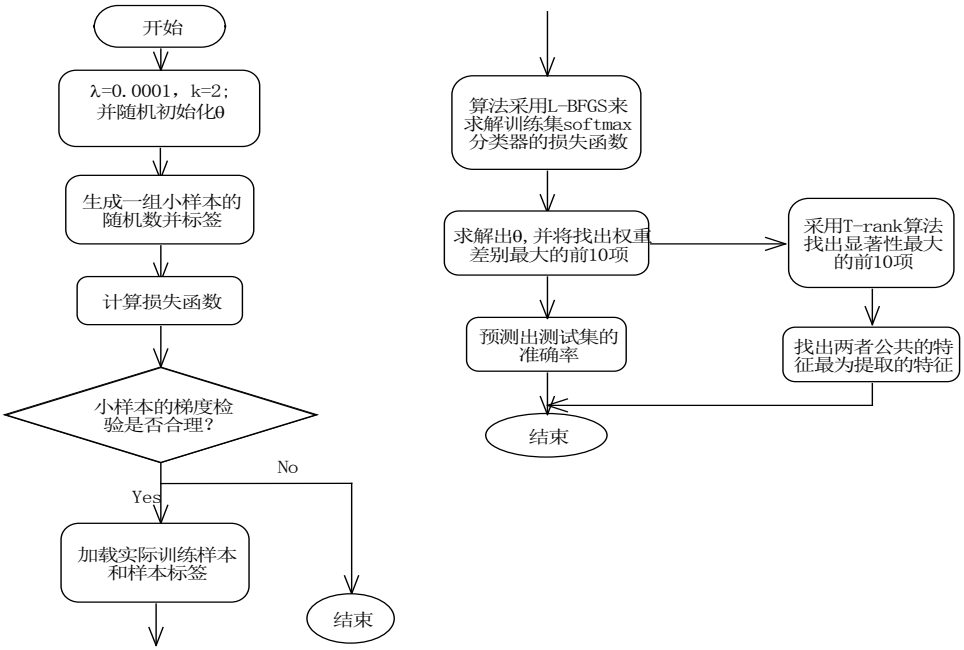

Figure 1. The flow chart by integrating T-rank and Softmax algorithm
图 1. 融合 T-rank 的 Softmax 算法流程图

3.3. 选用的分类器

为了说明特征提取的效果，我们采用常用的分类算法进行测试，本文以 BP 神经网络为例，进行特征提取后的分类准确率测试。

3.4. 实验方式及评价指标

在实际中可能会出现样本不平衡的情况，如样本总体为 100，其中 98 个是正常人，而只有 2 个是病人，如果算法判定所有样本均为正常人，则总的准确率仍可以达到 98%，而实际上算法对病人的并不具有很好的分类准确率，为此，我们定义以下分类准确率评价指标：

(1) 样本集总体的分类准确率： $Acc = \frac{TP + TN}{TP + FP + TN + FN}$。

(2) 样本集中正常人群体的分类准确率： $BAcc1 = \frac{TP}{TP + FP}$。

(3) 样本集中病人群体的分类准确率： $BAcc2 = \frac{TN}{TN + FN}$。

4. 实验结果分析

4.1. 特征提取结果

采用 Softmax 此方法后，就能得到高维特征量的权重(2*27,336 维)，然后分别对相对应的(正常人和患银屑病的人)权重进行做差取绝对值，然后将权重依次从大到小的排列，运行 20 次，每次都取权重为前 30 的特征基因，采用专家打分算法，最终选出前 10 个特征基因。

将 T 检验应用于基因表达数据即可检验基因在不同总体样本间表达差异是否显著。通过检验同一基因在正常人与银屑病患者之间的表达是否存在着显著性差异，可以判断该基因是否有可能是银屑病的致病基因之一。在此处，我们可以对每个基因分别进行 T 检验，取显著性最大，即正常人与银屑病患者表达差异最大的前十个基因作为特征值。将各个基因的 T 值(显著性)从小到大进行排序，取排名靠前的 10 个特征，并于 Softmax 算法得到的 10 个特征基因取交集，如图 2。

从图 2 也可以发现两种基因特征提取方法提取到的特征存在很大的重合性，从侧面说明了所提取的特征具有较强的稳健性。

4.2. 特征验证与结果分析

我们将数据分成 3 种不同方式进行训练和测试：

方式一：以 GSE14905 作为训练集，GSE13355 作为测试集；

方式二：以 GSE13355 作为训练集，GSE14905 作为测试集；

方式三：GSE14905 和 GSE13355 中各随机抽取一半作为测试和训练集

采用 BP 神经网络的算法，对本文提出的融合 T-rank 的 Softmax 的特征提取算法得到的 8 个基因进行分类准确性验证，验证结果如下表 1，表 2，表 3。

从表 1 中可知，当 GSE14905 作为训练集，GSE13355 作为测试集，其中有 22044(41469_at)，3280(203691_at)，27604(227736_at)这三个基因得到区分银屑病人的准确率为 100%。

从表 2 中可知，当以 GSE13355 作为训练集，GSE14905 作为测试集时，其中依然有 27604(227736_at)这个基因得到区分银屑病人的准确率为 100%。

从表 3 中可知，当以 GSE14905 和 GSE13355 中各随机抽取一半作为测试和训练集时，其中还有 9270(209720_s_at)，27604(227736_at)，9953(210413_x_at)这三个基因得到区分银屑病人的准确率为 100%。

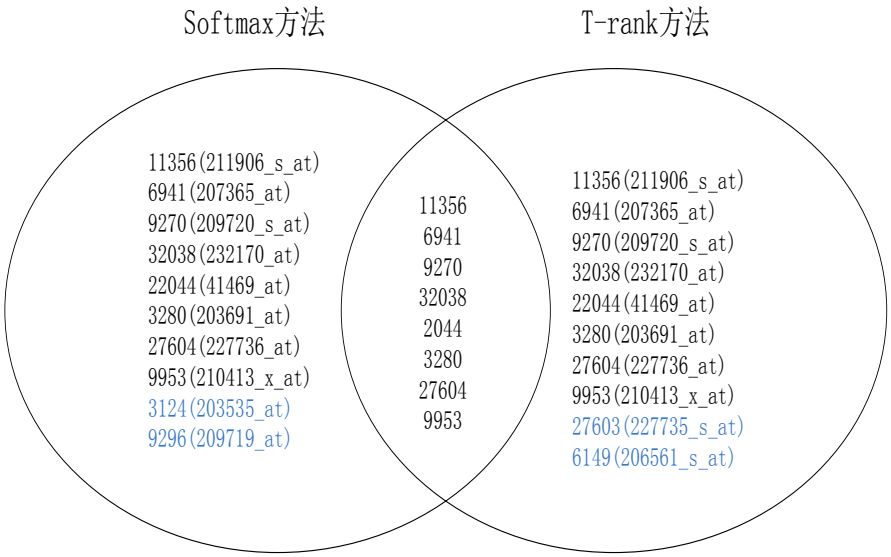

Figure 2. Information genes selected by Softmax algorithm and T-rank
图 2. Softmax 算法与 T-rank 法选择出的特征基因

Table 1. The Acc, BAcc1 and BAcc2 of 8 selected genes in case 1
表 1. 方式一 8 个基因编号及其对应的 Acc，BAcc1 及 BAcc2

编号(基因名称)	Acc	BAcc1	BAcc2
11356(211906_s_at)	97.54%	100%	97.03%
6941(207365_at)	100%	100%	100%
9270(209720_s_at)	98.36%	100%	98.02%
32038(232170_at)	88.52%	71.43%	92.08%
22044(41469_at)	100%	100%	100%
3280(203691_at)	100%	100%	100%
27604(227736_at)	100%	100%	100%
9953(210413_x_at)	97.54%	100%	97.03%

Table 2. The Acc, BAcc1 and BAcc2 of 8 selected genes in case 2
表 2. 方式二 8 个基因编号及其对应的 Acc，BAcc1 及 BAcc2

编号(基因名称)	Acc	BAcc1	BAcc2
11356(211906_s_at)	94.44%	95.24%	93.94%
6941(207365_at)	94.44%	95.24%	93.94%
9270(209720_s_at)	96.30%	95.24%	96.97%
32038(232170_at)	96.30%	100%	93.94%
22044(41469_at)	94.44%	95.24%	93.94%
3280(203691_at)	94.44%	95.24%	93.94%
27604(227736_at)	100%	100%	100%
9953(210413_x_at)	94.44%	95.24%	93.94%

Table 3. The Acc, BAcc1 and BAcc2 of 8 selected genes in case 3
表 3. 方式三 8 个基因编号及其对应的 Acc，BAcc1 及 BAcc2

编号(基因名称)	Acc	BAcc1	BAcc2
11356(211906_s_at)	96.59%	95.24%	97.01%
6941(207365_at)	97.73%	90.48%	100%
9270(209720_s_at)	100%	100%	100%
32038(232170_at)	98.86%	95.24%	100%
22044(41469_at)	98.86%	95.24%	100%
3280(203691_at)	97.73%	90.48%	100%
27604(227736_at)	100%	100%	100%
9953(210413_x_at)	100%	100%	100%

Table 4. The Acc, BAcc of 8 selected genes in comprehensive evaluation
表 4. 综合评价所得 8 个基因的 Acc 及 BAcc

编号(基因名称)	Average_Acc	Average_BAcc
11356(211906_s_at)	96.19%	96.41%
6941(207365_at)	97.39%	96.61%
9270(209720_s_at)	98.22%	98.37%
32038(232170_at)	94.56%	92.12%
22044(41469_at)	97.77%	97.40%
3280(203691_at)	97.39%	96.61%
27604(227736_at)	100%	100%
9953(210413_x_at)	97.33%	97.70%

最终我们综合考虑三种检测方式下各个特征基因对应不同群体的分类准确率，对测试的 8 个特征基因分别求出其在各个检测方式下的准确率的平均值，即

$$Average\_Acc = \frac{1}{3}\sum_{i=1}^{3} Acc\left(method\left(i\right)\right)$$

$$Average\_BAcc = \frac{1}{6}\left[\sum_{i=1}^{3} BAcc1\left(method\left(i\right)\right) + \sum_{i=1}^{3} BAcc2\left(method\left(i\right)\right)\right]$$

得到下述结果。

通过对表 4 的观察我们发现 27604(227736_at)基因的平均分类准确率达到了 100%。考虑到样本个数较少，为了保证模型的稳健性，同时我们也可以看到第 9270(209720_s_at)，22044(41469_at)，9953(210413_x_at)，这三个平均准确率次高的基因平均准确率也达到了 97%以上，具有很大的参考价值。因此可以取 27604(227736_at)，9270(209720_s_at)，22044(41469_at)，9953(210413_x_at)这四个基因作为最终的诊断基因。

为了便于对比，我们首先将 GSE14905 的合并数据 GSE13355 采用传统的 PCA (Principle Component Analysis)方法进行降维并取前 8 个主成分(累积贡献率达 81.3%)标准化得分来代替原始数据，然后利用 LDA (Linear Discriminant Analysis)方法对 8 个主成分同样在三种情况下进行分类验证，并计算出三种方

式下的平均值准确率 Average_Acc = 92.37%，Average_Bacc = 92.89%，可以看到本文算法提出的特征准确率明显高于 PCA 方法，更重要的是，PCA 虽然也能够有效的经行降维，但其得到的是所有特征的加权综合，无法判断出究竟是哪个具体基因才是关键诊断基因，这在实际的临床诊断操作上没有本文的方法方便实用。

5. 结论

本文通过融合 T-rank 的 Softmax 的特征提取算法，极大程度的利用了 Softmax 算法和 T-rank 算法的优点，与传统的 PCA+LDA 算法相比，有着极大的优势。该算法成功解决了低样本高维数特征提取的难题，而且能成功有效的提取关键特征基因，获得较高准确率疾病诊断模型。

基金项目

本文获教育部人文社科青年基金 (13YJC790105)、湖北工业大学博士科研启动基金(BSQD13050)、湖北工业大学基金项目(2015SW0204)资助。

参考文献 (References)

[1] 邹晶, 高磊, 李晋, 戴静珠, 李霞. 针对不同特征基因挖掘方法的特征基因功能一致性分析[J]. 中国生物医学工程学报, 2010, 29(2): 212-213.

[2] 李霞, 张田文, 郭政. 一种基于递归分类树的集成特征基因选择方法[J]. 计算机学报, 2004, 27(5): 675-682.

[3] 李颖新, 阮晓钢. 基于支持向量机的肿瘤分类特征基因基因选取[J]. 计算机研究与发展, 2005, 42(10): 1796-1801.

[4] 吕飒丽, 汪强虎, 李霞, 郭政. 基于决策森林特征基因的两种识别方法[J]. 生物信息学, 2004, 2(3): 19-22.

[5] 张飞, 王世祥, 王玲, 宋凯. 肺鳞状细胞癌癌症发展模式识别分类模型及特征基因识[J]. 生物化学与生物物理进展, 2016, 43(1): 63-74.

[6] Villasenor-Park, J., Wheeler, D. and Grandinetti, L. (2012) Psoriasis: Evolving Treatment for a Complex Disease. *Cleveland Clinic Journal of Medicine*, **79**, 413-423. http://dx.doi.org/10.3949/ccjm.79a.11133

[7] Yao, Y., *et al*. (2008) Type I Interferon: Potential Therapeutic Target for Psoriasis? *PLoS ONE*, **3**, e2737. http://dx.doi.org/10.1371/journal.pone.0002737

[8] Swindell, W.R., *et al*. (2011) Genome-Wide Expression Profiling of Five Mouse Models Identifies Similarities and Differences with Human Psoriasis. *PLoS ONE*, **6**, e18266. http://dx.doi.org/10.1371/journal.pone.0018266

[9] Nair, R.P., *et al*. (2009) Genome-Wide Scan Reveals Association of Psoriasis with IL-23 and NF-KappaB Pathways. *Nature Genetics*, **41**, 199-204. http://dx.doi.org/10.1038/ng.311

[10] Barrett, T., *et al*. (2011) NCBI GEO: Archive for Functional Genomics Data Sets—10 Years on. *Nucleic Acids Research*, **39**, D1005-D1010. http://dx.doi.org/10.1093/nar/gkq1184

[11] Irizarry, R.A., *et al*. (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, **4**, 249-264. http://dx.doi.org/10.1093/biostatistics/4.2.249

[12] Benito, M., *et al*. (2004) Adjustment of Systematic Microarray Data Biases. *Bioinformatics*, **20**, 105-114. http://dx.doi.org/10.1093/bioinformatics/btg385