引用格式:王晓,李辉,翟云清. 基于集成学习和基因本体标注库的细胞凋亡蛋白亚细胞位置预测[J]. 轻工学报,2016,31(4):95 – 101.
中图分类号:TP273 **文献标识码**:A
DOI:10.3969/j.issn.2096 – 1553.2016.4.014
文章编号:2096 – 1553(2016)04 – 0095 – 07

基于集成学习和基因本体标注库的细胞凋亡蛋白亚细胞位置预测

Predicting subcellular localization of apoptosis protein based on ensemble learning and Gene Ontology annotation database

王晓,李辉,翟云清
WANG Xiao,LI Hui,ZHAI Yun-qing

郑州轻工业学院 计算机与通信工程学院,河南 郑州 450001
College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, *China*

关键词:
亚细胞定位;同源蛋白;基因本体特征;集成 K 近邻算法

Key words:
subcellular location; homologous proteins; Gene Ontology(GO) features; ensemble K nearest-neighbor algorithm

摘要:针对目前凋亡蛋白的亚细胞定位预测精度不高的问题,提出了基于集成学习和基因本体(GO)标注库的细胞凋亡蛋白亚细胞位置预测方法. 该方法采用凋亡蛋白及其同源蛋白的 GO 特征,结合两层集成策略,预测凋亡蛋白的亚细胞位置. 在第一层,依据不同同源蛋白个数生成多个特征向量集合,选取距离权重 K 近邻分类器作为个体分类器,训练多个子预测模型,并以多数投票的方式集成. 在第二层,将第一层的集成模型作为子预测模型,以多数投票的方式集成不同近邻个数预测模型. Jackknife 检验结果表明:该方法在 CL317 凋亡蛋白数据集上预测准确率达到 96.2%,优于其他方法;此外,还有效降低了数据不均衡带来的影响.

收稿日期:2016 – 03 – 20
基金项目:国家自然科学基金项目(61402422);河南省教育厅科学技术研究重点项目(14A520063);郑州轻工业学院博士科研基金资助项目(2013BSJJ082)
作者简介:王晓(1982—),男,河南省卫辉市人,郑州轻工业学院讲师,博士,主要研究方向为机器学习与生物信息学.

Abstract: In order to deal with the problem that the prediction accuracy of subcellular localization of apoptosis proteins is not high, a method of predicting subcellular localization of apoptosis protein based on ensemble learning and Gene Ontology (GO) annotation database was proposed. This method utilized the GO features of apoptosis proteins and their homologous proteins combined with the two layer integration strategy to predict subcellular localization of apoptosis proteins. In the first layer, several sets of feature vectors were formulated by the different number of homologous proteins, then it selected the distance weighted *K*-nearest neighbor classifier as individual classifier, trained sub-prediction models, and integrated these models by majority voting. In the second layer, the prediction model of the first layer was used as the sub-prediction model, and it integrated the different nearest neighbors' sub-prediction models by the majority voting. The results of Jackknife test showed that prediction accuracy of the method reaches 96.2% on the CL317 apoptosis proteins dataset, which was superior to other methods. In addition, this method could reduce the impact of the data imbalance.

0　引言

细胞凋亡,也称为程序性细胞死亡,在许多生物存续过程中发挥着重要作用,比如破坏被感染细胞,从免疫系统消除自身反应性克隆等. 细胞的增殖与死亡能够使组织达到合适而稳定的细胞数,使机体达到动态的稳定. 细胞凋亡一旦出现问题,各种疾病则会随之而来,如癌症和自身免疫性疾病等,造成机体功能的紊乱[1-3]. 凋亡蛋白在程序性细胞死亡的机制中发挥着核心作用,了解凋亡蛋白质的功能,有助于理解这一机制[4]. 鉴于蛋白质的功能已被证明与其亚细胞位置紧密相关[5],因而获取有关凋亡蛋白亚细胞位置的信息,有助于理解凋亡蛋白功能与细胞凋亡机制.

生物学实验可以进行蛋白质的亚细胞定位,但手工实验的方法代价高、时间长. 随着新发现蛋白质的数量成指数级增长,依靠实验来注释蛋白质已远远不能满足需求,然而借助计算机自动化预测却能克服这些困难[6].

在过去的十几年,国内外研究者在预测凋亡蛋白亚细胞位置方面做过许多工作,2003 年,G. Zhou 等[7]首次提出凋亡蛋白亚细胞定位的问题,构建了包含 98 个细胞凋亡蛋白,4 种亚细胞位置的 ZD98 数据集,采用了基于氨基酸组分(AAC)的协变判别式算法; A. Bulashevska 等[8]采用贝叶斯分类方法进行定位预测;Z. H. Zhang 等[9]通过构建 ZW225 数据集,采用分组质量编码结合支持向量机(EBGW_SVM)进行定位预测;Y. L. Chen 等[10-11]构建了一个新的 CL317 数据集,含 317 凋亡蛋白,6 个亚细胞位置,使用离散增量(ID)和 SVM 预测凋亡蛋白质亚细胞位置;Y. Ding 等[12]采用伪氨基酸组分(PseAAC)和模糊 *K* 近邻算法(FKNN)进行预测;L. Zhang 等[13]使用距离频率结合 SVM 算法进行定位预测;J. Qiu 等[14]使用小波系数改进预测性能;T. G. Liu 等[15]使用基于位置特异性的记分矩阵的自动协方差转换(PSSM-AC)的方法提高预测性能;H. Lin 等[16]使用 PseAAC 和 SVM 进一步提高预测精度;Q. Gu 等[17]采用集成学习和特征选择的方法构建预测模型;X. Yu等[18]采用基于氨基酸置换矩阵的自动协方差转换法改善预测效果;V. Saravanan等[19]使用自适应集成进一步提高预测效果;T. Liu 等[20]使用基于 PSSM 的 tri-gram 编码方法提取特征,取得了极大的性能提升. 这些研究存在两个共性问题:其一,仅仅使用氨基酸的序列特性来描述蛋白质,而忽略了一些隐藏于蛋白质内的重要特征,比如,蛋白质结构信息、蛋白质功能信息等;其二,预测凋亡蛋白亚细胞位置时,使用的算法多是机器学习传统算法,其本身具有一定的局限性,预测效果一般. 鉴于此,本文拟提出

基于集成学习和基因本体 GO(Gene Ontology)标注库的细胞凋亡蛋白亚细胞位置预测方法,以期提高凋亡蛋白定位预测精度.

1 本文方法

1.1 特征提取

随着蛋白质注释工作的发展和不断深入,出现了很多蛋白质注释数据库,它们对蛋白质的功能进行标注与描述,其内容还能随着研究的不断深入而更新. 与蛋白质的序列特征相比,蛋白质注释信息能包含更多的蛋白质特征信息,从而显著提升预测性能. GO 数据库是众多蛋白质注释数据库中的一种,它使用 GO 语义来描述蛋白质的分子和生物功能特性. 这些语义分为 3 种不同的种类:细胞组件、分子功能和生物过程[21].

基因本体联合会同其他生物数据库合作完成 GO 术语与基因产物之间的联系,对它们所包含的基因产物使用 GO 的定义方法进行注释[22-23],即每个基因或基因产物都会有与之相关的 GO 术语列表. 在 GOA 数据库中,1 个 GO 语义可能与多个蛋白质的访问号相关联,1 个访问号可能与多个 GO 语义相对应.

GO 术语的数量随着时间的推移在迅速增加,不可能将所有的 GO 术语用来生成特征向量,否则可能导致高维灾难等问题. 本文的方法表述如下:

1)GO 标号的压缩和重组. 选取 GO 数据库(2015-06-23)中只标注是"细胞组件"的 GO 标号,其中包含 3 951 个标注为"细胞组件"的 GO 标号. 由于 GO 标号不是连续的,所以要对其重新进行排序,原来的 GO 编号 GO:0000015, GO:0000108, GO:0000109, …, GO:1990777,将分别变为 GO_compress:0001, GO_compress:0002, GO_compress:0003, …, GO_compress:3951. 原始的 GO 数据库在经过处理之后获得新的 GO 数据库,称为 GO_compress 数据库.

2)蛋白质表示为

$$\boldsymbol{P}=[f_1 \quad f_2 \quad f_3 \cdots f_u \cdots f_{3951}]^{\mathrm{T}}$$

其中

$$f_u=\frac{\sum_{j=1}^{N_{\boldsymbol{P}}^h}\theta(u,j)}{N_{\boldsymbol{P}}^h} \quad u=1,2,3\cdots,3\ 951$$

式中,$N_{\boldsymbol{P}}^h$ 表示 $\boldsymbol{P}$ 和 $\boldsymbol{P}$ 的同源蛋白的数量;如果第 j 个蛋白质的 GO 标号集合中包含第 u 个GO 标号,$\theta(u,j)=1$,否则为 0.

3)使用 BLAST 工具(版本 2.2.30)搜索 SWISS-PROT 数据库(2015-07-24)找到蛋白质 $\boldsymbol{P}$ 的同源蛋白集.

4)若 $\boldsymbol{P}$ 没有同源蛋白质,也就是说 $\boldsymbol{P}$ 同源蛋白集为空,则只使用 $\boldsymbol{P}$ 自身搜索 GO 数据库,生成与 $\boldsymbol{P}$ 相对应的 GO 标号集合,将其映射到 GO_compress 数据库中,并在相应位置置 1,否则置 0. 例如蛋白质 $\boldsymbol{P}$ 的 GO 标号集合中包含 GO:0000108,对应为 GO_compress:0002,即 $f_2=1$. 若 $\boldsymbol{P}$ 同源蛋白集不为空,则搜索 GO 数据库找到 $\boldsymbol{P}$ 同源蛋白的 GO 标号,也将其映射到 GO_compress 数据库中.

1.2 基于集成学习的预测算法

传统 K 近邻(KNN)算法的基本思想是:根据设定的 K 值,从训练样本中选择出与测试样本距离最近的 K 个样本,然后统计这 K 个样本中分别属于每一类的样本数,将待测样本归入样本数最多的类别. 该算法有很明显的不足,当各类样本数量不平衡时,如一个类的样本数量很大,而其他类的样本很小时,可能导致在预测新样本时,该样本的 K 个邻居中所属大容量类的样本占多数,而出现分类错误的情况. 为了避免出现这一问题,本文采用基于距离权重的 KNN 分类器,计算两个样本之间距离,权值与样本之间的距离成反比:距离越小权重越大,将权重累加,最后选择累加值最高的类作为该待

测样本的类.

分类器的性能与 K 值是密切相关的,设置 $K=\{3,4,\cdots,8\}$. 依据前文所描述的特征提取方法提取数据集中每一条凋亡蛋白的特征,对于每一条凋亡蛋白选取 λ 条同源蛋白,$\lambda=\{1,2,\cdots,10\}$. 如果一条凋亡蛋白没有这么多的同源蛋白质,则选取其所有同源蛋白质. 使用 $NN(\lambda,K)$ 表示在这两个参数上训练的子预测器,就得到了一个分类器的集合:

$$\{NN(\lambda,K)\}=\begin{Bmatrix} NN(1,3) & NN(2,3) & \cdots & NN(10,3) \\ NN(1,4) & NN(2,4) & \cdots & NN(10,4) \\ \vdots & \vdots & \ddots & \vdots \\ NN(1,8) & NN(2,8) & \cdots & NN(10,8) \end{Bmatrix}$$

其中,$NN(1,3)$ 表示选取一条同源蛋白来提取特征,同时选取 KNN 预测算法的近邻数目 K 为 3,由此训练得到的子分类器,以此类推.

本文采用两层集成策略,通过多数投票策略构成第一层集成分类器为

$$Q(K)=NN(1,K)\oplus NN(2,K)\oplus NN(3,K)\oplus\cdots\oplus NN(10,K)$$

其中,$\oplus$ 表示集成符号,K 值保持不变. 集成分类器 $Q(K)$ 预测结果定义为

$$B_K^t=\begin{cases}1 & Score(K)\geqslant 0.5\\ 0 & Score(K)<0.5\end{cases}$$

其中,$Score(K)=\dfrac{\sum_{i=1}^{10}b_i^t}{10}$,$b_i^t$ 表示子预测器 $NN(i,K)$ 对亚细胞位置 $t(t=1,2,\cdots,6)$ 的预测输出,如果预测器预测蛋白质 $\boldsymbol{P}$ 属于亚细胞位置 t,则 $b_i^t=1$,否则 $b_i^t=0$.

集成分类器 $Q(K)$ 通过多数投票策略构成的第二层集成分类器为

$$R=Q(3)\oplus Q(4)\oplus\cdots\oplus Q(8)$$

其预测输出定义为

$$C_t=\frac{\sum_{K=3}^{8}B_K^t}{6}$$

则预测未知位置蛋白质 $\boldsymbol{P}$ 的亚细胞位置为最大值 C_t 所对应的亚细胞位置.

1.3　预测流程

输入一条蛋白质序列 $\boldsymbol{P}$,先使用 BLAST 工具搜索 Swiss-Prot 数据库找到 $\boldsymbol{P}$ 的同源蛋白质集合,如果没有同源蛋白质,则只用蛋白质 $\boldsymbol{P}$ 本身,然后再搜索 GO 数据库找到 $\boldsymbol{P}$ 及其同源蛋白质的 GO 标号集合,生成特征向量,输入分类器,最后查看结果. 预测流程图如图 1 所示.

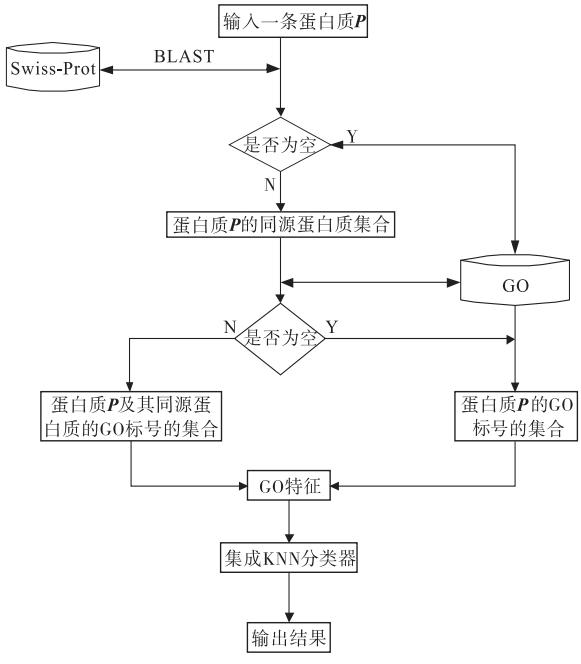

图 1　预测过程的流程图

Fig. 1　A flowchart to show the prediction process

2　实验结果与讨论

2.1　数据集与评价指标

使用 CL317 数据集[10]作为独立数据集,在数据集中包含了 6 类亚细胞,其中有 112 个细胞质蛋白(Cytoplasmic Proteins)、55 个膜蛋白(Membrane Proteins)、34 个线粒体蛋白(Mitochondrial Proteins)、17 个分泌蛋白(Secreted Proteins)、52 个细胞核蛋白(Nuclear Proteins)、47 个内质网蛋白(Endoplasmic Reticulum Proteins). 在细胞质蛋白集合中“P03405”和

“Q07814”这两个凋亡蛋白在 Uniprot 数据库(2015 – 07 – 24)中已替换为“P03404”和“Q07812”,而在细胞质蛋白集合中已经包含这两个凋亡蛋白,细胞质蛋白数量变更为 110 个,细胞核蛋白“Q9Z1S4”已于 2009 年 11 月 3 日从数据库中删除,细胞核蛋白的数量为 51 个,更新后 CL317 蛋白质总数为 314 个,每个亚细胞位置上包含的蛋白数量见表 1.

表 1 每个亚细胞位置上包含的蛋白数量

Table 1 Number of proteins in each subcellular location

编号	亚细胞位置	蛋白质数量
1	细胞质(Cy)	110
2	细胞膜(Me)	55
3	线粒体(Mi)	34
4	分泌 (Se)	17
5	细胞核(Nu)	51
6	内质网(En)	47

常用的检验方法主要有 3 种:独立检验、*k*-fold 交叉检验和 Jackknife 检验. 其中,Jackknife 检验被认为是最严格和客观的检验方法之一. 在 Jackknife 检验中,将 N 个蛋白质的数据集分为 N 个互不相交的子集,也就是将每一条蛋白质都作为一个子集依次取出作为测试,其余 $N-1$ 个蛋白质作为训练集,循环 N 次,每次抽取的样本都要放回数据集. 本文采用 Jackknife 检验方法,选择敏感度 *SN*,特异性 *SP*,马氏相关系数 *MCC*,总体预测正确率 *ACC* 作为评价性能的指标.

$$ACC=\frac{\sum_i TP_i}{N}\quad SN_i=\frac{TP_i}{TP_i+FN_i}\quad SP_i=\frac{TN_i}{TN_i+FP_i}$$

$$MCC=\frac{TP_iTN_i-FP_iFN_i}{\sqrt{(TP_i+FP_i)(TP_i+FN_i)(TN_i+FP_i)(TN_i+FN_i)}}$$

2.2 实验结果

表 2 展示了数据集中蛋白质经过 Jackknife 检验的预测结果. 由表 2 可见,本文方法在各个评价指标上的预测精度都超过了 95%,取得了非常好的预测效果. 此外,从各个亚细胞位置的 *MCC* 指标值可以看出,本文方法能够有效降低数据不均衡带来的影响.

不同的方法在 CL317 数据集上预测结果见表 3. 从表 3 可以看出,本文方法所取得总体预测精度 *ACC* 要高于其他方法. 一些亚细胞位置如 Me,Mi,En,它们的敏感度 *SN* 值要高于其他方法,特别是对于 En,其所有蛋白均能正确预测,这是其他方法所不能实现的. 然而我们也意识到 Cy 和 Nu 的敏感度 *SN* 要比 APSLAP 略低,这可能是因为一些凋亡蛋白及其同源蛋白的 GO 注释信息较少,对于一个蛋白质,较为丰富的 GO 特征信息可以提高预测准确率.

3 结语

本文使用凋亡蛋白及其同源蛋白 GO 注释信息替代传统的氨基酸序列信息来描述蛋白质. 因为蛋白质的 GO 注释比氨基酸序列能够包含更多必要的信息,故能显著提升凋亡蛋白亚细胞位置的预测性能. 凋亡蛋白质与其同源蛋白本身具有一些相同特性,加入同源蛋白的 GO 特征信息,可弥补蛋白质本身 GO 特征信息的不足. 在预测算法方面,使用两层集成的策略,能进一步提升预测效果. 实验结果表明本文方法在凋亡蛋白亚细胞位置预测方面是非常有效的. 为了更好地服务生物学研究,下一步计划把本文所描述的方法开发成在线预测平台.

表 2 数据集中蛋白质经 Jackknife 检验的预测结果

Table 2 The prediction result for the data set by Jackknife

位置	*SN*/%	*SP*/%	*MCC*	*ACC*/%
Cy	97.3	97.0	0.937	96.2
Me	98.2	100.0	0.989	
Mi	97.1	98.2	0.907	
Se	94.1	100.0	0.968	
Nu	88.2	99.6	0.916	
En	100.0	100.0	1.000	

表 3　不同的方法在 CL317 数据集上预测结果

Table 3　Comparison of different methods on CL317 data set　%

方法	*SN*						*ACC*
	Cy	Me	Mi	Se	Nu	En	
ID[10]	81.3	81.8	85.3	88.2	82.7	83.0	82.7
ID_SVM[11]	91.1	89.1	79.4	58.8	73.1	87.2	84.2
DF_SVM[13]	92.9	85.5	76.5	76.5	93.6	86.5	88.0
Auto_Cova[18]	86.4	90.7	93.8	85.7	92.1	93.8	90.0
FKNN[12]	93.8	92.7	82.4	76.5	90.4	93.6	90.9
PseAAC_SVM[16]	93.8	90.9	85.3	76.5	90.4	95.7	91.1
EN_FKNN[17]	98.2	83.6	79.4	82.4	90.4	97.9	91.5
APSLAP[19]	99.1	89.1	85.3	88.2	84.3	95.8	92.4
PSSM_tri-gram[20]	98.2	96.4	94.1	82.4	96.2	95.7	95.9
本文方法	97.3	98.2	97.1	94.1	88.2	100.0	96.2

参考文献:

[1] EVAN G, LITTLEWOOD T. A matter of life and cell death[J]. Science, 1998, 281(5381): 1317.

[2] REED J C, PATERNOSTRO G. Postmitochondrial regulation of apoptosis during heart failure [J]. Proc Natl Acad Sci USA, 1999, 96(14): 7614.

[3] JACOBSON M D, WEIL M, RAFF M C. Programmed cell death in animal development[J]. Cell, 1997, 88(3): 347.

[4] SCHULZ J B, WELLER M, MOSKOWITZ M A. Caspases as treatment targets in stroke and neurodegenerative diseases[J]. Annals of Neurology, 1999, 45(4): 421.

[5] SUZUKI M, YOULE R J. Structure of Bax: Coregulation of dimer formation and intracellular localization[J]. Cell, 2000, 103(4): 645.

[6] 张松, 黄波, 夏学峰, 等. 蛋白质亚细胞定位的生物信息学研究[J]. 生物化学与生物物理进展, 2007(6): 573.

[7] ZHOU G P, Doctor K. Subcellular location prediction of apoptosis proteins[J]. Proteins: Structure, Function and Genetics, 2003, 50(1): 44.

[8] BULASHEVSKA A, EILS R. Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains[J]. BMC Bioinformatics, 2006, 7(1): 298.

[9] ZHANG Z H, WANG Z H, ZHANG Z R, et al. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine[J]. FEBS Letters, 2006, 580(26): 6169.

[10] CHEN Y L, LI Q Z. Prediction of the subcellular location of apoptosis proteins[J]. Journal of Theoretical Biology, 2007, 245(4): 775.

[11] CHEN Y L, ZHONG Q Z. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition[J]. Journal of Theoretical Biology, 2007, 248(2): 377.

[12] DING Y, ZHANG T. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier[J]. Pattern Recognition Letters, 2008, 29(13): 1887.

[13] ZHANG L, LIAO B, LI D, et al. A novel repre-

sentation for apoptosis protein subcellular localization prediction using support vector machine [J]. Journal of Theoretical Biology, 2009, 259 (2): 361.

[14] QIU J, LUO S, HUANG J, et al. Predicting subcellular location of apoptosis proteins based on wavelet transform and support vector machine [J]. Amino Acids, 2010, 38(4): 1201.

[15] LIU T G, ZHENG X Q, WANG J, et al. Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation [J]. Protein & Peptide Letters, 2010, 17(10): 1263.

[16] LIN H, WANG H, DING H, et al. Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition [J]. Acta Biotheoretica, 2009, 57(3): 321.

[17] GU Q, DING Y, JIANG X, et al. Prediction of subcellular location apoptosis proteins with ensemble classifier and feature selection [J]. Amino Acids, 2010, 38(4): 975.

[18] YU X, ZHENG X, LIU T, et al. Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation [J]. Amino Acids, 2012, 42(5): 1619.

[19] SARAVANAN V, LAKSHMI P T V. APSLAP: an adaptive boosting technique for predicting subcellular localization of apoptosis protein [J]. Acta Biotheoretica, 2013, 61(4): 481.

[20] LIU T, TAO P, LI X, et al. Prediction of subcellular location of apoptosis proteins combining tri-gram encoding based on PSSM and recursive feature elimination [J]. Journal of Theoretical Biology, 2015, 366: 8.

[21] HARRIS M A, CLARK J, IRELAND A, et al. The Gene Ontology (GO) database and informatics resource [J]. Nucleic Acids Research, 2004, 32(Database issue): D258.

[22] CAMON E, MAGRANE M, BARRELL D, et al. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology [J]. Nucleic Acids Research, 2004, 32 (Database issue): D262.

[23] CAMON E, MAGRANE M, BARRELL D, et al. The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro [J]. Genome Research, 2003, 13(4): 662.