基于拟比对 CNN 方法的人类 p53 癌症基因二级数据库构建及分析

王丹丹， 李晨鸿， 徐海阳， 蔡 蓉， 朱 平*

（江南大学 理学院，江苏 无锡 214122）

摘要： 以 NCBI 维护的一级数据库为数据源建立人类癌症 p53 核苷酸序列二级数据库，该数据库设计主要包括 4 个方面：癌症信息、p53 序列信息、样本信息和参考文献信息。以 XML 格式为中间格式保存一级数据库数据，并通过解析提交到二级数据库，初步实现数据的检索、链接和统计分析等功能。本文提出一种拟比对 CNN 方法对p53 癌症基因序列进行比对分析，通过改善传统 CNN 相似度评估公式，增强两序列全局比对相似度的敏感性和可靠性。结果表明，将改进的序列比对算法应用于乳腺癌和非小细胞肺癌p53 外显子基因序列比对，发现外显子 5 突变后序列比对结果存在较大差异，可以作为区别这两种癌症的参考。此外，通过将一级数据库以 XML 形式转化成二级数据库，实现了网络数据与本地数据的动态交换。

关键词： 二级数据库；癌症；p53 基因序列；细胞神经网络；序列比对

中图分类号：Q 811.4　**文章编号：**1673-1689(2019)04-0015-06　**DOI：**10.3969/j.issn. 1673-1689.2019.04.003

Construction and Analysis of Secondary Database of Human Cancer Gene p53 Based on Quasi Alignment Cellular Neural Network

WANG Dandan， *LI Chenhong*， *XU Haiyang*， *CAI Rong*， *ZHU Ping**

（School of Science，Jiangnan University，Wuxi 214122，China）

Abstract： Using the biological primary databases at the National Center for Biotechnology Information （NCBI），We construct a secondary database of human cancer-related nucleotides p53. The database design mainly includes four aspects：cancer information，p53 sequence information，sample information and reference information. We store the data from NCBI in XML file，then by parsing the files to secondary database and initially realize the data of searching，linking ，statistical analysis and other functions. p53 cancer gene sequences are compared by quasi alignment one dimensional cellular neural network method，and the sensitivity and reliability of the global alignment of the two sequences are enhanced by improving the similarity evaluation formula. We

收稿日期： 2016-07-01

基金项目： 国家自然科学基金项目(11271163)。

* **通信作者：** 朱　平(1962—)，女，博士，教授，主要从事数据挖掘，计算分子生物学，理论计算机科学等方面的研究。

E-mail：zhuping@jiangnan.edu.cn

引用本文： 王丹丹，李晨鸿，徐海阳，等. 基于拟比对 CNN 方法的人类 p53 癌症基因二级数据库构建及分析[J]. 食品与生物技术学报，2019，38(04)：15-20.

applied the improved sequence alignment algorithm in non small cell lung cancer and breast cancer p53 gene sequence alignment , the result shows that there are great differences in mutant p53 Exon 5 of two cancer sequences which can be used to discriminate these two cancers. In addition, by transforming the primary database into the secondary database in the form of XML, the dynamic exchange of network data and local data is redized, which provides an excellent platform for the study of cancer and p53 gene.

Keywords: secondary database,cancer ,p53 gene sequence ,cellular neural network ,sequence alignment

随着各种模式生物基因组计划的相继完成或全面实施,有关核酸、蛋白质的序列和结构等生物学数据呈指数增长,越来越多基因的结构和功能得到阐明,建立简洁、专用性强和数据质量高的二级数据库及分析系统已成为研究热点之一[1-2]。

各类二级数据库的建立是研究生物信息学的重要出发点。一般而言,生物信息一级数据库的数据来源于原始实验室的直接提交,而对一级数据库的数据信息进行搜索、加工和整理成二级数据库,已广泛应用于生物信息学、生命科学等领域[3-5]。目前被公认是癌症生物标志物的 p53 抑癌基因最初被认为是一种癌基因,随着近十年研究的深入,p53 作为抑癌基因的功能逐渐被揭示,对 p53 基因的研究也日益受到重视[6-7]。其中汪小霞等[8]揭示了 EZH2 与 p53 表达的正相关性以及 EZH2 蛋白在肿瘤中受 p53 基因调控的现象;MAKWANE 等[9]通过聚合酶链反应单链构象多态性分析发现人类乳腺癌 p53 肿瘤抑制基因的外显子 5 和 7 存在突变。p53 基因突变与人类肿瘤的关系十分密切,已成为肿瘤发病的主要因素,而随着生物医学的发展,很多癌症可以进行基因治疗, 因此, 有必要建立一个关于 p53 基因的二级生物信息数据库对 p53 基因进行深入分析研究。

在生物信息学中,基因序列比对已成为一种处理基因信息的基本方法,有利于发现生物序列的功能、结构和进化信息。目前,序列比对算法主要包括 Smith-Waterman 算法、BLAST 算法和 FASTA 算法等[10-11]。Smith-Waterman 算法敏感性强,但复杂度很高,运算速度较慢,后 2 种序列比对算法具有更高的运算速度,但敏感性较差。而一维细胞神经网络(cellular neural network,CNN)[12-13] 序列比对算法是一种新提出的序列比对算法,具有较低时间复杂度和良好的敏感性。本研究采用拟比对一维 CNN 序列比对算法应用于不同癌症 p53 基因序列比对,并改善算法相似度评估公式,大大增强了序列比对结果的敏感性和可靠性。

本研究中主要针对 p53 癌症基因的外显子和内含子数据,构建一个二级生物信息数据库。通过 NCBI(National Center for Biotechnology Information)一级数据库收集 p53 癌症基因序列,经过解析与归纳将二级数据库设计为包含 p53 信息、癌症信息、样本信息和参考文献信息等 4 个部分,实现二级数据库的信息查寻与使用, 并通过 Agent 程序实现数据库的自动更新和维护。采用改进相似度评估公式的 CNN 序列比对算法,增强了序列比对敏感性,使乳腺癌和非小细胞肺癌 p53 外显子基因序列比对结果更可靠。

1 p53 癌症基因二级数据库构建

1.1 数据源

NCBI 是国际主要生命科学信息服务机构之一,每天都有大量来自实验室和测序机构发布的序列数据进入该数据库,并保持与其他数据库的数据交换和更新,因而汇集了当前所有公开的核酸和蛋白质序列, 本二级数据库的数据主要来源于 NCBI 维护的基因数据库。

本研究中采用 p53、human、cancer、exon、intron 等关键词搜索一级数据库,通过代理程序自动获取数据库 Web 信息资源,并对其进行检索、归纳和转换产生二级数据库,其流程图如图 1 所示。

本二级数据库目前已经收集了包括乳腺癌、非小细胞肺癌、胃癌、肝癌等 16 种癌症的 516 条 p53 癌症基因序列。

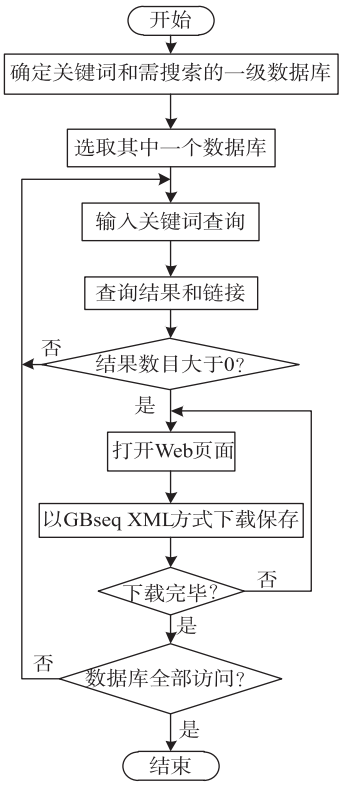

图 1　下载一级数据库流程

Fig. 1　Download the flow chart of primary database

1.2　数据解析

本研究所构建的二级数据库主要过程见图 2。首先从 NCBI 一级数据库中下载 p53 癌症基因,并保存为网页信息,然后使用 XML Document 类 DOM 模型创建、修改、遍历 XML 文档,运行后台解析程序对 XML 文档进行映射,实现批量导入数据。最后,采用 XML 技术将各种异构数据源数据转换成 XML 公共数据模型格式,实现网络数据资源和二级数据库的数据交换,同时以 GBseq XML 格式获取文本数据并构建本地二级数据库。

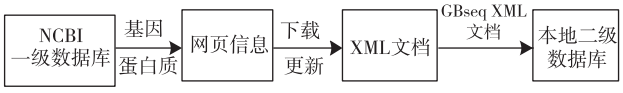

图 2　构建二级数据库的系统结构

Fig. 2　System architecture of the construction of secondary database

1.3　数据结构设计

1.3.1　数据库结构　本研究所构建的二级数据库主页包含了 1 个集合,这个集合包含各种癌症子集合,每个子集合包含癌症 p53 基因序列的外显子或内含子信息。同时以图片形式展示人体各部分癌症位点,用户可以点击图片进行浏览,也可通过关键字搜索查找 p53 基因序列编号,如果数据库中有这一条序列则会自动跳转到对应序列,如未能找到相应 p53 序列编号,则不显示任何信息。

当用户点击某种癌症图片进入对应页面,可以看到包括 p53 信息、癌症信息、样本信息、参考文献信息等实体信息。用户如对某个癌症 p53 基因序列的详细信息感兴趣,可以点击 more 按键详细查看 NCBI 一级数据库中这条序列的完整信息,同时也可链接到 p53 研究的原始文献进一步阅读,查看包括癌症名称、p53 外显子或内含子、样本信息(样本数、样本来源)以及参考文献信息(题目、发表年、PMID)等具体信息。

本二级数据库的实体信息主要包括以下 4 个部分:(1) p53 信息,主要包含 p53 基因序列在数据库中编号 (Accession)、p53 某个外显子或者内含子基因序列以及外显子对应的蛋白质序列、每条序列的长度和起始子位置;(2) 癌症信息,主要包含癌症的名称、癌症的分类;(3) 样本信息,主要包含样本来源、样本数、样本类型、样本的研究方法;(4) 参考文献信息,主要包含文献题目、PMID、发表日期、备注等。

1.3.2　数据库功能模块　本数据库系统分为 5 个模块:第 1 个模块为癌症 p53 基因外显子或内含子序列显示系统,包含了各种癌症及癌症部位显示图,用户可以通过点击癌症图片的方式,了解具体癌症 p53 基因序列信息。

第 2 个模块为数据库介绍模块,介绍了数据库的基本内容,包括数据库更新信息、数据统计信息、数据内容介绍。

第 3 个模块为搜索系统,用户可以通过输入癌症 p53 序列编号,了解该 p53 序列是否被测序出来,是否在数据库中,如果有,则数据库会给出相应的序列信息,用户可通过点击的方式了解详细信息。

第 4 个模块为用户数据提交系统,作为一个大的整合型的数据库,我们尽可能收集所有数据,如果用户有新数据可以提交至我们的数据库,将由后台人员进行数据审查。

第 5 个模块为数据共享,即作为一个公开的数据库,数据是共享的,数据库中的信息是以 Excel 表形式进行汇总,全面涵盖了所存储数据,用户可以对数据进行下载使用。

1.4 二级数据库的管理与更新

二级数据库的管理系统能够有效减轻管理员的负担,目前已经有多种不同环境下的数据库管理系统。二级数据库管理主要步骤包括增删相应序列和编辑二级数据库注释信息、二级数据库更新需与一级数据库更新同步，主要包括修改序列条目、删除冗余条目和加入新条目。

随着一级数据库中已有数据的不断变更以及新数据的不断加入,人工更新方式已难以满足二级数据库的实时有效更新,因此,需要采用一种自动获取 Web 信息的方法实现二级数据库自动更新。而 SQL Server 提供的企业级管理软件能够进行常规任务的自动化管理,管理员可以通过 SQL Server 提供的 Agent 服务实现数据库的自动管理和更新。利用 Agent 服务在特定日期及时检测一级数据库中的版本信息，二级数据库通过接收 Agent 服务消息并与二级数据库对应的版本号进行比较,主动对二级数据库进行更新,若版本号变动则自动下载更新该条目,通过匹配一级数据库中有关 p53 基因信息条目的版本号,自动判断该条目是否已在二级数据库中存在,不存在则通过文本消息通知管理员添加该条目[3,14]。与人工更新方式相比,SQL Server Agent 服务能够高效地完成数据库自动管理和更新。

2 拟比对 CNN 序列比对方法

一维 CNN 模型具有一个线性细胞排列结构,每个细胞最多有两个相连细胞,这些特点可以有效地来进行两条 DNA 序列比对。一维 CNN 与传统 CNN 不同,因为它仅由两个单独的一维细胞神经网络组成,一个固定的主子网络,一个可移动的从子网络,分别代表两条 DNA 序列片段,其中网络中每一个细胞都对应 DNA 序列中的碱基。算法中从子网络随时间以固定距离移动,计算主子网络中每一个细胞在不同时刻的状态值,最终将所得到的状态值进行排列形成状态矩阵,通过动态规划方法产生一条全局比对最优路径。在这条路径的引导下,通过在合适的位置插空使得两条长度不同的 DNA 序列变成长度相同新的 DNA 序列，然后对这两条 DNA 序列片段进行全局比对。

一维 CNN 计算公式如下:

$$x_{1,i}(t+1)=\left(1-\frac{1}{CR_x}\right)x_{1,i}(t)+\frac{1}{C}\left[\sum_{C_i(k)\in N_{1,i}(r,t)}A_k y_{l,k}(t)+\sum_{C_i(k)\in N_{1,i}(r,t)}B_k u_{l,k}+I_{1,i}\right]$$

$$y_{1,i}(t)=f[x_{1,i}(t),I_{1,i}]\begin{cases}1 & \text{if } x_{1,i}(t)=i_{1,j}\\ 0 & \text{else}\end{cases} \quad (1)$$

其中,$x_{1,i}(t)$表示主子网络中细胞 i 在 t 时刻的状态,$y_{1,i}(t)$表示从细胞 i 中接收到的反馈输出,A_k,B_k 分别是反馈模板 A 和控制模板 B 的系数,I_i,R_x 和 C 是 3 个常量,$y_{1,k}(t)$和 $u_{1,i}(t)$分别为细胞 k 在 t 时刻的输出和相关输入。

具体序列比对步骤如下:

Step1:设置 CNN 初始值,将 DNA 序列化为可计算的数字特征,最基本的特征{*,A,C,G,T}相应的量化为{0,−1,−0.5,0.5,1},“*”代表空格。

Step2:通过状态选择函数计算状态矩阵,其中状态选择函数计算公式为:

$$x_{1,i}(t)=\begin{cases}x_{1,i}(t) & \text{if } i=0 \text{ and } t=1\\ \max(\alpha_1,\alpha_2,\alpha_3) & \text{else}\end{cases} \quad (2)$$

其中参数 α_1,α_2,α_3 满足以下公式:

$$\begin{cases}\alpha_1=x_{1,i-1}(t-1)-F_3;\\ \alpha_2=x_{1,i-1}(t-2)+y_{1,i}(t)\times(F_1+F_2)-F_2;\\ \alpha_3=x_{1,i}(t-1)-F_3\end{cases} \quad (3)$$

则状态矩阵为:

$$\boldsymbol{R}_{n\times m}=\begin{bmatrix}x_{1,0}(1) & x_{1,1}(2) & \cdots & x_{1,m-1}(m)\\ x_{1,0}(2) & x_{1,1}(3) & \cdots & x_{1,m-1}(m+1)\\ \vdots & \vdots & \ddots & \vdots\\ x_{1,0}(n) & x_{1,1}(n+1) & \cdots & x_{1,m-1}(m+n-1)\end{bmatrix} \quad (4)$$

Step3:根据状态矩阵形成全局比对的路径。通过最后的状态逐步回溯到第 1 个状态,其中每个状态选取规则为包含上 1 个状态左上角 2×2 矩阵中的最大值。例如,状态矩阵中的 m 行 n 列的状态为(n,m)则其下 1 个状态选取规则为在$(n,m-1)$、$(n-1,m-1)$、$(n-1,m)$中选取最大值,然后将其作为下 1 个状态值。

Step4：根据所选路径对 2 条 DNA 序列进行全局比对,如果前后 2 个状态横坐标相差 1,纵坐标相差 0,则在第一条序列中插入 1 个空格,如果前后 2 个状态横坐标相差 0,纵坐标相差 1,则在第 2 条序列中插入 1 个空格,其他情况则序列保持不变。

Step5：计算 2 条序列比对的相似度评估。将 CNN 序列比对方法中的计算相似度公式改进为:

$$\mathrm{Sim}(S_1,S_2)=\frac{1}{2}\left[\frac{N_{\mathrm{ma}}}{L(S_1)}+\frac{N_{\mathrm{ma}}}{L(S_2)}\right]\times 100\% \quad (5)$$

其中 N_{ma} 为步骤 4 中 2 条序列匹配的个数,$L(S_1)$、$L(S_2)$分别为序列 S_1、S_2 的长度。假设两序列 S_1、S_2 中 S_1 长度较短,可知 $N_{ma}\in\{0,1,\cdots\cdots,L(S_1)\}$,则由式(5)可以得出 $\mathrm{Sim}(S_1,S_2)\in\left[0,\frac{1}{2}+\frac{L(S_1)}{2\times L(S_2)}\right]$,当两序列匹配个数为零时,其最小相似度为零;当两序列匹配个数达到最大,即较短序列的长度时,其相似度最大,该最大相似度与两序列长度相关,当两序列长度相同时,最大相似度为 1,当两序列长度相差较大时,最大相似度趋近于 0.5。因此,改进后的相似度评估公式增强了序列比对敏感性,使得序列比对结果更可靠。

其中将公式(1)中的初始值取为 $C=1$,$R_x=1$,$I=0$,$A=\{0,0,0\}$,$B=\{0,1,-1\}$,$F=\{F_1,F_2,F_3\}=\{5,4,2\}$。

定义 1:称如上定义的 CNN 方法为拟比对 CNN 方法。

3 基于拟比对 CNN 的 p53 癌症基因序列比对

为了探讨不同癌症 p53 基因之间的区别和联系,本研究以乳腺癌和非小细胞肺癌为例进行序列比对,分析这两种癌症之间的联系与区别。分别对乳腺癌和非小细胞肺癌 p53 基因的 Exon4、Exon5、Exon6、Exon7、Exon8 和 Exon9 进行拟比对 CNN 算法比对,其中乳腺癌的 Exon4 有 15 条,Exon5 有 28 条,Exon6 有 27 条,Exon7 有 38 条,Exon8 有 39 条,Exon9 有 50 条;非小细胞肺癌的 Exon4 有 13 条,Exon5 有 25 条,Exon6 有 20 条,Exon7 有 35 条,Exon8 有 37 条,Exon9 有 48 条。

分别采用 CNN 方法、BLAST 序列比对方法和拟比对 CNN 方法对 2 种癌症 p53 基因同一外显子序列进行比对,表 1 中拟比对 CNN 方法利用步骤 5 所给出的两序列相似度评估公式(5)对 Exon4、Exon5、Exon6、Exon7、Exon8、Exon9 的 DNA 序列进行全局比对,其中 Mean 为平均值,Std 为标准差。

表 1 拟比对 CNN、CNN 与 BLAST 方法的全局相似度比较

Table 1 Global similarity comparisons between quasi alignment CNN、CNN and BLAST algorithms

方法		序列					
		Exon4	Exon5	Exon6	Exon7	Exon8	Exon9
BLAST	Mean	97.29%	51.35%	94.94%	91.17%	92.68%	90.82%
	Std	0.192 4	0.096 5	0.174 9	0.167 3	0.172 1	0.135 2
CNN	Mean	97.98%	52.61%	95.75%	93.03%	93.71%	93.37%
	Std	0.135 4	0.074 2	0.154 6	0.154 2	0.163 5	0.131 2
拟比对 CNN	Mean	99.75%	55.34%	97.32%	95.31%	95.27%	95.73%
	Std	0.102 4	0.065 4	0.144 8	0.130 1	0.128 7	0.095 4

从表 1 中可以看出,拟比对 CNN 方法相比 CNN 方法和 BLAST 序列比对方法有较高的全局相似度,尤其当序列长度较长时拟比对 CNN 相比 CNN 和 BLAST 全局相似度提高较多。此外,采用拟比对 CNN 方法对两种癌症 p53 基因的外显子序列比对相似度的标准差相对较小。表明拟比对 CNN 方法相比于其他两种方法不仅具有较好的敏感性,而且其结果的可靠性良好。

从表 1 还可以看出,采用拟比对 CNN 方法对乳腺癌和非小细胞肺癌的 p53 基因外显子进行比对,其结果有很大不同。其中 Exon5 相比于 Exon4、Exon6、Exon7、Exon8 和 Exon9 而言相似度最低,只有 55.34%,而 Exon4 的相似度最高,达到了 99.75%,其他几个外显子也都超过了 95%,这表明乳腺癌和非小细胞肺癌 p53 的 Exon5 序列突变存在较大差异。因此,在突变的 p53 外显子中 Exon5 可以作为区别乳腺癌和非小细胞肺癌参考标准,而其他几个 p53 癌症基因外显子难以作为区别乳腺癌和非小细胞肺癌的参考标准。

将拟比对 CNN 方法与 CNN 方法和 BLAST 序列比对方法对乳腺癌和肺癌的 Exon4、Exon5、Exon6、Exon7、Exon8、Exon9 序列比对时间进行比较,结果如表 2 所示。

由于拟比对 CNN 方法对 2 条 p53 基因序列进行比对,其时间复杂度为 $O(T)=O(m,n)\approx O(m+n+1)$,而 BLAST 比对算法的时间复杂度为 $O(T)=O(m\times n)$较高,其中 m、n 分别为 2 条序列的长度,所以拟比对 CNN 方法有效降低了算法的时间复杂度。

从表 2 可以看出,就计算时间而言不论是较长的序列还是较短的序列拟比对 CNN 方法和 CNN 方法都比 BLAST 所需要的计算时间短很多,且就拟比对 CNN 方法本身而言序列总长度越长比对所需计算时间也越长。因此,采用拟比对 CNN 方法对基因

表 2 拟比对 CNN、CNN 与 BLAST 方法的计算时间比较

Table 2 Computation time comparisons of the quasi alignment CNN、CNN and BLAST algorithms ms

方法	序列					
	Exon 4	Exon 5	Exon 6	Exon 7	Exon 8	Exon9
BLAST	131.42	66.37	90.13	134.16	166.86	157.43
CNN	102.55	57.69	64.63	108.09	128.58	141.90
拟比对 CNN	101.35	56.24	64.57	108.93	127.26	141.45

序列进行比对有效提高了算法运算效率。

4 结 语

本研究以癌症生物标志物 p53 基因为对象,通过 NCBI 一级数据库收集 p53 癌症基因序列, 经过解析与归纳将二级数据库设计为包含 p53 信息、癌症信息、样本信息和参考文献信息等 4 个部分,利用 5 个功能模块更加有利于实现二级数据库的信息查询与使用。为了进一步提高 CNN 序列比对方法相似度的敏感性和可靠性,本文对相似度评估公式进行改进,并将其定义为拟比对 CNN 方法,有效提高了算法的性能。采用拟比对 CNN 方法对乳腺癌和非小细胞肺癌 p53 基因的外显子序列进行比对分析,发现两种癌症 p53 基因的 Exon5 序列突变存在较大差异,可作为区别乳腺癌和非小细胞肺癌的参考标准。

后续将增加更多癌症 p53 基因,以及加入某种癌症的其他生物标志物比如乳腺癌的 HER2、ER,结直肠癌的 EGFR、KRAS 等,使数据库的内容更全面,实用性更强,应用性更广。

参考文献:

[1] CHEN Jian,HU Ruihua,YANG Keyang,et al. The preliminary construction of mice strains resources bioinformatics secondary database based on NET[J]. **Laboratory Animal Science**,2011,28(6):43-47.(in Chinese)

[2] DU J,CAO X Q. Construction of molecular biology secondary database resources platform [J]. **Hans Journal of computational Biology**,2014,4:32-41.

[3] CHENG Peng,HUANG Zhigang,HONG Yahui,et al. Construction and application of a secondary database for Phytohormone-related nucleotides and proteins[J]. **Chinese Bulletin of Botany**,2010,46(2):258-264.(in Chinese)

[4] TEUFEL A. Bioinformatics and database resources in hepatology[J]. **Journal of Hepatology**,2015,62(3):712-719.

[5] XU Sangang,ZHUANG Yonglong,HAO Zhiyong,et al. The construction of pathogenic plasmodium molecular functional annotation secondary database[J]. **Chinese Journal of Biomedical Engineering**,2012,31(6):882-888.(in Chinese)

[6] MARK W,LI Yaocheng,GEOFFREY M W. MDM2,MDMX and p53 in oncogenesis and cancer therapy [J]. **Nature Reviews Cancer**,2013,13(2):83-96.

[7] SUN T Z,CUI J. Dynamics of p53 in response to DNA damage:mathematical modeling and perspective [J]. **Progress in Biophysics and Molecular Biology**,2015,119:175-182.

[8] WANG Xiaoxia,MENG Gang,LI Li,et al. Expression of EZH2 and p53 in breast cancer and their clinical significance [J]. **Chinese Journal of Clinical and Experimental Pathology**,2015,31(3):273-276.(in Chinese)

[9] MAKWANE N,SAXENA A. Study of mutations in p53 tumour suppressor gene in human sporadic breast cancer [J]. **Indian Journal of Clinical Biochemistry**,2009,24(3):223-228.

[10] RALF S N,SURENDRA K,KAMRAN S T. Blast output visualization in the new sequencing era [J]. **Briefings in Bioinformatics**,2014,15(4):484-503.

[11] PEI Songwen,WANG Xinyi,WEI Gang,et al. Research on parallel BLAST algorithm based on multi-core stream processors[J]. **Journal of System Simulation**,2011,23(10):2065-2069.(in Chinese)

[12] JI L P,PU X R,QU H. One-dimensional pairwise CNN for the global alignment of two DNA sequences [J]. **Neurocomputing**,2015,149:505-514.

[13] RAHIMEH R,MEHDI J,SHOHREH K,et al. Benign and malignant breast tumors classification based on region growing and CNN segmentation[J]. **Expert Systems with Applications**,2015,42(3):990-1002.

[14] YAO Yangchun,CHAI Shiyou,LU Xing,et al. New generation of distributed object-oriented real-time database management system[J]. **Power System Technology**,2007,31(2):284-287.(in Chinese)