Journal of Computer Applications
计算机应用,2014,34(11):3117-3120
ISSN 1001-9081
CODEN JYIIDU
2014-11-10
http://www.joca.cn

文章编号:1001-9081(2014)11-3117-04
doi:10.11772/j.issn.1001-9081.2014.11.3117

基于消息传递接口的大规模生物网络比对并行化算法

束俊辉[1],张　武[1,2],薛倩斐[1],谢　江[1,2*]
(1.上海大学 计算机工程与科学学院,上海 200444;　2.上海大学 高性能计算中心,上海 200444)
(*通信作者电子邮箱 jiangx@shu.edu.cn)

摘　要:为有效降低生物网络比对算法的时间复杂度,提出一种基于可扩展的蛋白质相互作用网络比对(SPINAL)算法的消息传递接口(MPI)并行化实现方法。该方法将MPI并行化思想运用在SPINAL算法中,在多核环境中采用并行排序代替算法原本的排序方式,并结合负载均衡策略合理分配任务。实验结果表明,与未使用并行排序以及负载均衡策略相比,该方法在处理大规模生物网络比对时能有效地缩短计算时间,提高运算效率,对于不同组比对数据都有较为稳定的优化保障,具有良好的可扩展性。

关键词:生物网络比对;并行网络比对;可扩展的蛋白质相互作用网络比对;并行排序;消息传递接口
中图分类号:TP391;TP311　　**文献标志码**:A

Parallel alignment algorithm of large scale biological networks based on message passing interface

SHU Junhui[1], ZHANG Wu[1,2], XUE Qianfei[1], XIE Jiang[1,2*]
(1. *School of Computer Engineering and Science, Shanghai University, Shanghai* 200444, *China*;
2. *High Performance Computing Center, Shanghai University, Shanghai* 200444, *China*)

Abstract: In order to reduce the time complexity of biological networks alignment, an implementation for large scale biological networks alignment based on Scalable Protein Interaction Network Alignment (SPINAL) in Message Passing Interface (MPI) program was proposed. Based on MPI, the SPINAL algorithm combined with parallelization method was applied into this approach. Instead of serial algorithm, parallel sorting algorithm was used in multi-core environment. Load balancing strategy was chosen to assign tasks reasonably. In the processing of large scale biological networks alignment, the experiment shows that, compared with the algorithm without parallelization and load balancing strategy, this proposed algorithm can reduce the runtime and improve computation efficiency effectively.

Key words: biological networks alignment; parallel networks alignment; Scalable Protein Interaction Network Alignment (SPINAL); parallel sorting; Message Passing Interface (MPI)

0　引言

生物网络比对对于寻找生物间相似功能域、预测生物结构功能相关性以及发现物种之间的进化关系等有着重要的意义和作用,是当今生物信息学研究重点之一。其中,针对蛋白质相互作用(Protein-Protein Interaction, PPI)网络的研究是该领域的热点问题。Singh等[1]使用特征函数约束的方法,通过网络中相邻节点的相似度函数计算对应节点的相似度得到相似网络;Memisevic等[2]研究的算法使用贪心策略比对得到匹配边数最大的网络;El-Kebir等[3]利用拉格朗日松弛算法对稀疏生物网络比对进行研究;Xie等[4]将反馈机制运用于网络比对研究并将其理论应用于帕金森病相关网络;可扩展的蛋白质相互作用网络比对(Scalable Protein Interaction Network Alignment,SPINAL)算法[5]是一种启发式算法,该算法将比对图问题转化成最大权重二分图问题进行求解。然而,随着生物学的快速发展和生物网络数据量的迅速增长,这些算法都面临着同一瓶颈,即运行时间过长。为了缩短算法运行时间,提高计算效率,并行计算无疑是解决这一问题的良好方法。综合对现有算法的考量,本文选择无论从结果稳定性还是比对评价指标方面都有一定优势的SPINAL算法,通过消息传递接口(Message Passing Interface,MPI)结合并行排序、负载均衡策略实现了SPINAL并行化算法。

1　生物网络比对

所谓生物网络比对,即找到生物网络节点之间的映射关系,使得网络之间的相似性得分最高[6]。本文主要针对生物网络中的蛋白质相互作用网络进行研究。结合文献[5-9]给出一些蛋白质相互作用网络比对相似性得分相关概念的形式化定义。

设PPI网络 $G_1=(V_1,E_1)$,$G_2=(V_2,E_2)$,其中 V_1、V_2 代表 G_1、G_2 的蛋白质节点集合,E_1 与 E_2 代表 G_1、G_2 边的集合。设生物网络比对匹配结果为 $A_{12}=(V_{12},E_{12})$,其中 V_{12} 代表匹

收稿日期:2014-06-05;**修回日期**:2014-08-20。　**基金项目**:国家自然科学基金资助项目(91330116);高等学校博士学科点专项科研基金资助项目(20113108120022);上海市科学技术委员会重点科研项目(11510500300)。

作者简介:束俊辉(1990-),男,上海人,硕士研究生,主要研究方向:生物信息学、高性能计算;　张武(1957-),男,江西武宁人,教授,博士,CCF会员,主要研究方向:高性能计算、生物信息学、计算流体学;　薛倩斐(1989-),女,浙江建德人,硕士研究生,主要研究方向:生物信息学、高性能计算;　谢江(1971-),女,湖北恩施人,副教授,博士,CCF会员,主要研究方向:生物信息学、高性能计算。

配节点对集合$\{\langle x_i, y_j\rangle \mid x_i \in V_1, y_j \in V_2\}$,$V_{12}$ 中节点对$\langle x_i, y_j\rangle$需满足以下条件:对于任意的$\langle x_i, y_j\rangle \in V_{12}$、$\langle x_i', y_j'\rangle \in V_{12}$,$x_i \neq x_i'$ 且 $y_i = y_j'$;E_{12} 代表匹配边集合,当$(\langle x_i, y_j\rangle, \langle x_i', y_j'\rangle) \in E_{12}$、$\langle x_i, y_j\rangle \in V_{12}$、$\langle x_i', y_j'\rangle \in V_{12}$时,$(x_i, x_i') \in E_1$ 且$(y_j, y_j') \in E_2$。

G_1 与 G_2 的匹配边数 $|E_{12}|$ 代表两个蛋白质相互作用网络拓扑结构的相似性,匹配节点 V_{12} 间的相似性意味着两个 PPI 网络生物学意义的相似性,因此匹配边数以及匹配节点相似值对于判断两个 PPI 网络是否相似起重要作用。在此根据文献[5]给出蛋白质相互作用网络相似性得分计算公式:

$$S(A_{12}) = \alpha \times |E_{12}| + (1-\alpha) \times \sum_{\forall \langle x_i, y_j\rangle \in V_{12}} sim(x_i, y_j) \tag{1}$$

其中:$sim(x_i, y_j) \in [0,1]$;$x_i \in V_1$;$y_j \in V_2$;$\alpha \in [0,1]$;$sim(x_i, y_j)$ 代表两个蛋白质节点的相似值,具体相似值可根据具实际问题求解。本文中的 $sim(x_i, y_j)$ 是通过序列比对查询工具[10](Basic Local Alignment Search Tool,BLAST)的期望值归一化得出。平衡因子 α 可调节节点相似性以及边匹配值的权重,可结合实际蛋白质相互作用网络进行调优,使生物网络相似性得分 S 达到理想值。

2 SPINAL 算法及其并行化实现

2.1 SPINAL 算法

SPINAL 算法是一个启发式蛋白质相互作用网络比对算法,其目标是使式(1)中的相似性得分 S 达到最大值。与其他 PPI 网络比对算法相比,其在 PPI 网络比对评价指标:相似性得分以及基因本体一致性(Gene Ontology Consistency,GOC)得分方面都有较为明显的优势,因此本文选择 SPINAL 算法作为研究对象,对其进行并行化。以下是 SPINAL 算法执行步骤:

第 1 步　输入 PPI 网络 G_1、G_2,相似值函数 sim,平衡因子 α。

第 2 步　循环遍历所有节点 x_i、y_j,其中 $x_i \in V_1$,$y_j \in V_2$。初始化置信得分矩阵 $\boldsymbol{P}(x_i, y_j)$,其计算公式如式(2):

$$\boldsymbol{P}(x_i, y_j) = \alpha \times \frac{|deg_{G_1}(x_i) - deg_{G_2}(y_j)|}{|\max(deg(G_1)) - \max(deg(G_2))|} + (1-\alpha) \times sim(x_i, y_j) \tag{2}$$

其中:$deg_{G_1}(x_i)$、$deg_{G_2}(y_j)$ 分别为 x_i、y_j 在 G_1、G_2 中的度;$\max(deg(G_1))$、$\max(deg(G_2))$ 分别代表为 G_1、G_2 中的最大度。

第 3 步　将置信得分矩阵 $\boldsymbol{P}$ 复制至 $\boldsymbol{P}'$,循环遍历所有节点 $x_i \in V_1$,$y_j \in V_2$,构造邻居二分图 $NBG(\{\langle x_i, y_j\rangle\}, \boldsymbol{P}')$,根据式(3)求出 $\boldsymbol{P}(x_i, y_j)$。

$$\boldsymbol{P}(x_i, y_j) = \alpha \times \frac{\sum_{\langle x_i, y_j\rangle \in C} \frac{\boldsymbol{P}'(x_i, y_j)}{deg_{G_1}(x_i) \times deg_{G_2}(y_j)}}{\sqrt{|C|}} + (1-\alpha) \times sim(x_i, y_j) \tag{3}$$

以下是本步骤中涉及到的邻居二分图定义:邻居二分图 $NBG(\{\langle x_i, y_j\rangle\}, \boldsymbol{F})$ 是完全二分图,$\{\langle x_i', y_j'\rangle\}$ 构成二分图节点对,$x_i' \in N(x_i)$、$y_j' \in N(y_j)$,$N(x_i)$、$N(y_j)$ 分别代表着 x_i, y_j 邻居节点集合;二分图的权值由节点对$\langle x_i, y_j\rangle$的权重映射关系 $\boldsymbol{F}(x_i, y_j)$ 求得。贡献者(contributor)集合 C 定义为基于贪心算法的 $NBG(\{\langle x_i, y_j\rangle\}, \boldsymbol{P})$ 的最大权重匹配集合。

第 4 步　将 $\boldsymbol{P}'$ 复制至 $\boldsymbol{P}$,判断置信得分矩阵 $\boldsymbol{P}$ 是否收敛,若未收敛执行第 3 步;收敛则执行第 5 步。

第 5 步　构建栈 D,将$\langle x_i, y_j\rangle$收敛求得的 $\boldsymbol{P}(x_i, y_j)$ 值按从小到大的顺序插入栈 D 中。

第 6 步　弹出栈顶元素$\langle x_i, y_j\rangle$。

第 7 步　构造$\langle x_i, y_j\rangle$的 $NBG(\{\langle x_i, y_j\rangle\}, \boldsymbol{P})$ 以及贡献者集合 C,将 C 中的所有$\langle u_i, v_j\rangle$插入 V_{12}。

第 8 步　若栈 D 中存在元素则执行第 6 步;如栈中无元素,输出结果,程序结束。

2.2 SPINAL 算法并行化

SPINAL 算法虽然在评价指标方面有较好的结果,但是当其处理数据规模庞大的蛋白质相互作用网络时,需耗费较长的计算时间。为提高其计算效率,本文基于 MPI 对 SPINAL 算法进行并行化,结合了 MPI 良好的并行性和可移植性的优点,最大限度地减少算法运行时间,提高计算效率。

在算法各执行步骤中,第 3 步构建 P 矩阵的时间复杂度为 $O(k|V_1||V_2|d_1d_2\log(d_1d_2))$,其中:$k$ 为迭代次数;d_1、d_2 为 G_1、G_2 中的最大度。它的时间复杂度相对于算法其他步骤是最高的。因此,本文对 SPINAL 算法进行并行化的工作主要集中在 $\boldsymbol{P}$ 矩阵构建上。算法的基本并行化方法描述如下:

1)主进程根据输入值,初始化置信得分矩阵 $\boldsymbol{P}$。

2)主进程将 $\boldsymbol{P}$ 矩阵通过 MPI 广播至其他各个进程。

3)将所需计算的 $\boldsymbol{P}$ 矩阵划分成 N 块,N 块分配给 N 个进程,各进程按划分到的任务进行计算。

4)MPI 归约各进程计算结果。

5)判断 $\boldsymbol{P}$ 矩阵是否收敛。若收敛,则执行 6);若不收敛,则执行 2)。

6)根据 $\boldsymbol{P}$ 矩阵计算得到匹配结果。

以上为 SPINAL 算法的基本并行化方法。但是通过实验发现,基于基本并行方法的 SPINAL 算法会出现各进程排序耗时所占比重较大以及负载不均衡问题,具体实验结果见 3.3 节。为改善上述问题,本文提出并行优化方法:并行排序策略和负载平衡策略。

2.2.1 并行排序策略

各进程在迭代计算置信得分矩阵的过程中,需要对贡献者集合 C 进行构建。构建 C 中所需的阈值是对当前置信得分矩阵各元素值进行排序后筛选得到的。由于数据规模庞大,各进程对置信得分矩阵排序需耗费大量时间,然而在每次迭代过程中,各进程的排序结果是相同的。所以,本文采用了并行排序的策略,以减少各进程进行串行排序所花费的时间。

本文采用经典的并行采样排序[11](Parallel Sorting by Regular Sampling,PSRS)对排序进行并行化。具体步骤如下:首先,将 $\boldsymbol{P}$ 矩阵中的所有元素按总进程数 n 进行平均划分,各进程根据分配到的数据进行快速排序并且抽取 $n-1$ 个样本发送到主进程。其次,主进程接收到各进程发来的抽样样本后进行归并排序,根据归并排序结果抽样选取 $n-1$ 个分裂点样本,并将这些样本发送至其他各个进程。再者,各进程收到后对已排序的数据按照分裂点进行分段,将 $k(k<n)$ 段发送给第 k 个进程,各进程根据接收到的数据进行归并排序,并在归并排序完成后将数据发送至主进程。最后,主进程接受所有发来的数据,将数据归并排序后,即得到最终排序的结果,并将

最后排序的结果广播至其他进程。

通过以上并行排序方法,不仅充分发挥 MPI 多进程并行优势,而且将一个进程单独完成的排序工作分配给多个进程并行完成,从而明显减少各进程每轮迭代过程中所花费在排序过程中的时间。

2.2.2　负载平衡策略

在并行计算中,各进程任务划分对于并行计算运行效果的优劣有着非常大的影响作用。在 SPINAL 并行化过程中,任务划分主要集中在各进程计算置信得分矩阵 ***P*** 上。而计算 ***P*** 矩阵的运算量是根据比对的两个 PPI 网络节点的度来决定的。由于 PPI 网络中各蛋白质节点的度分布不均,如果只是使用普通的按行划分的方式进行任务分配,很可能会导致各个进程运算量的负载不均。为了减少因为负载不均而造成的运算时间延长的问题,本文使用负载平衡策略对计算 ***P*** 矩阵任务划分方式进行优化,如图 1 所示。

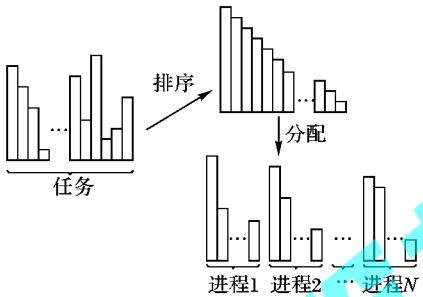

图 1　负载平衡策略

以下为具体任务划分方式:

1) 按行划分需要计算的置信得分矩阵 ***P***,得到任务集合,然后根据 ***P*** 矩阵行号所对应节点的度的大小将集合中的任务进行排序。

2) 按照进程号先正序后逆序的周期对排序后的任务进行分配。

通过上述步骤能使每个进程分配到较平均的任务量,从而达到负载平衡的目的。

根据以上描述,本文综合了并行化基本策略、并行化排序以及负载平衡优化策略,实现 SPINAL 生物网络比对并行化算法,算法流程如图 2 所示。

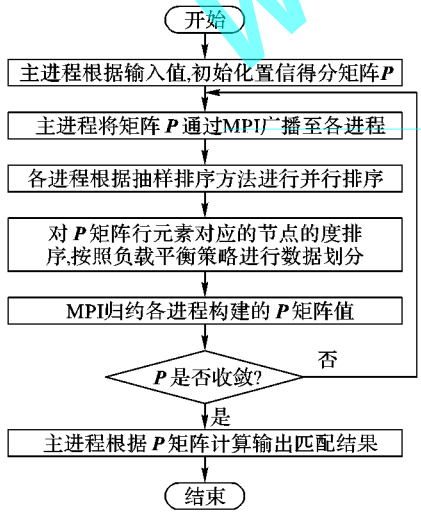

图 2　SPINAL 并行算法流程

在整个算法中,涉及进程间通信的有主进程广播、归约矩阵 ***P***、并行排序三个部分。其中,主进程广播、归约矩阵 ***P*** 只需在每次迭代过程中各进行一次通信,而在并行排序过程中,虽然进程间需要进行多次通信,但是仅仅传递节点度数集合,数据规模较小。由此可见,以上通信时间与构建 ***P*** 矩阵所需的计算时间相比,在整个算法中所占比重较小。因此,算法并行的研究重点集中在构建 ***P*** 矩阵上。

3　实验与分析

3.1　实验环境

该算法是在上海大学高性能计算集群自强 4000 上实现的。实验使用 2 台计算节点,每台 2 颗 Intel SandyBridge 架构的 E5-2690 CPU(2.9 GHz/8-core),64 GB 内存,节点间通过 56 Gb/s FDR Infiniband连接。实验使用 Centos 6.3 操作系统,C++ 为编程语言,MPI 库版本为 MPICH2。

3.2　实验数据

实验数据选取了人类、果蝇、酵母三个 PPI 网络数据集,所有数据集来源于 ISOBASE[12] 数据库。各数据集详细说明见表 1。

表 1　PPI 网络数据集

数据集	节点数	边数
果蝇	7518	25635
人类	9633	34327
酵母	5499	31261

3.3　结果与分析

3.3.1　未使用负载平衡与使用负载平衡策略结果对比

为了验证本文负载平衡策略的有效性,在此对未使用负载平衡策略与使用负载平衡策略的各进程运行时间进行对比。本文选用数据规模较大的酵母和人类蛋白质网络比对进行实验,使用 8 个进程进行比对运算。图 3 是未使用负载平衡各进程运行时间对比图,从中可看出各进程运行时间分布不均,运算时间最大值与最小值差别较为明显。图 4 是使用负载平衡策略的各进程运行时间对比图。从图 4 可看出经过优化后,各进程的运算时间相对较为平均,运行时间最大值与最小值差别不大。可见使用负载平衡优化策略对各节点任务平均分配起到了较好的效果。

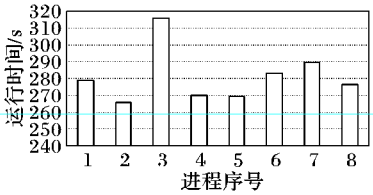

图 3　未使用负载平衡策略进程运行时间

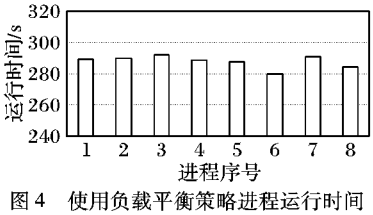

图 4　使用负载平衡策略进程运行时间

3.3.2　并行优化结果

本节对 SPINAL 算法采用并行优化方法(使用并行排序以及负载均衡策略)与未使用并行优化方法的算法进行对比

实验。选用数据规模较大的酵母和人类蛋白质网络比对进行实验,2 ~ 32 个进程进行比对运算。实验结果如图 5 所示,可看出随着进程数的增加,并行优化算法与未优化算法的加速比差距明显增大。当进程数达到 32 时,前者的加速比是后者的 2.7 倍。可见实现并行优化策略是非常有必要的。

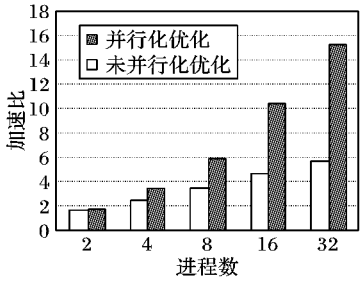

图 5 并行优化效果对比

3.3.3 大规模蛋白质网络比对

对于并行算法,加速比和运行时间是衡量算法优劣的关键因素。在此使用了酵母和人类、酵母和果蝇、人类和果蝇的三组 PPI 网络比对进行实验。实验中使用了 2 到 32 不等的进程数。图 6 展示的是 SPINAL 并行算法加速比,图 7 是算法的运行时间。从图 6 中可知,随着进程数的增加,三组网络比对运行加速比都呈线性增长,达到了良好的加速效果。从图 7可看出,随着进程数的增加,各 PPI 网络比对时间不断减少。其中运行时间最长的酵母和人类网络比对,运行时间从 1 619 s 减少到 106 s。通过并行化大大降低了比对运行耗时,提高了计算效率。另外根据图 6 ~ 7 中的实验结果还可以发现,尽管数据集不同,并行算法运行结果仍然能够保持一个较为稳定的加速比增加并且运行时间减少的趋势。从这点可以得出,本文算法对于不同组比对数据都有较为稳定的优化保障,具有良好的可扩展性。

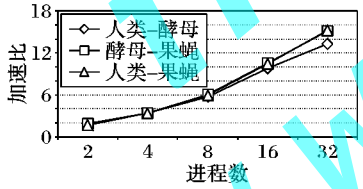

图 6 SPINAL 并行化算法加速比

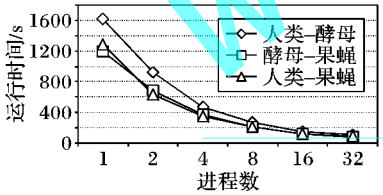

图 7 SPINAL 并行化算法运行时间

4 结语

本文提出一种基于 SPINAL 生物网络比对算法的 MPI 并行化实现,在基本并行化思想上采用并行化排序以及负载平衡策略优化方法对 SPINAL 算法进行并行化,较好地缩短了算法运行时间,提高了大规模生物网络数据比对的计算效率,有效地解决了传统生物网络比对算法在面临大规模数据运算耗时的瓶颈。

参考文献:

[1] SINGH R, XU J, BERGER B. Global alignment of multiple protein interaction networks with application to functional orthology detection [J]. Proceedings of the National Academy of Sciences, 2008, 105(35): 12763 - 12768.

[2] MEMISEVIC V, PRZULJ N. C-GRAAL: common-neighbors-based global graph alignment of biological networks[J]. Integrative Biology, 2012, 4(7): 734 - 743.

[3] EL-KEBIR M, HERINGA J, KLAU G W. Lagrangian relaxation applied to sparse global network alignment[C]// PRIB 2011: Proceedings of the 6th IAPR International Conference on Pattern Recognition in Bioinformatics, LNCS 7036. Berlin: Springer-Verlag, 2011: 225 - 236.

[4] XIE J, ZHANG S, WEN T, *et al*. A querying method with feedback mechanism for protein interaction network[C]// Proceedings of the 2011 First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology. Piscataway: IEEE Press, 2011: 351 - 358.

[5] ALADAQ A E, ERTEN C. SPINAL: scalable protein interaction network alignment[J]. Bioinformatics, 2013, 29(7): 917 - 924.

[6] GUO XL, GAO L, CHEN X. Models and algorithms for alignment of biological networks[J]. Journal of Software, 2010, 21(9): 2089 - 2106. (郭杏莉, 高琳, 陈新. 生物网络比对的模型与算法[J]. 软件学报, 2010, 21(9): 2089 - 2106.)

[7] CHINDELEVITCH L, LIAO C S, BERGER B. Local optimization for global alignment of protein interaction networks [EB/OL]. [2013-10-10]. http://psb.stanford.edu/psb-online/proceedings/psb10/chindelevitch.pdf.

[8] KUCHAIEV O, PRSULJ N. Integrative network alignment reveals large regions of global network similarity in yeast and human[J]. Bioinformatics, 2011, 27(10): 1390 - 1396.

[9] ZASLAVSKIY M, BACH F, VERT J P. Global alignment of protein-protein interaction networks by graph matching methods[J]. Bioinformatics, 2009, 25(12): i259 - 1267.

[10] ALTSCHUL S F, GISH W, MILLER W, *et al*. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403 - 410.

[11] SHI H, SCHAEFFER J. Parallel sorting by regular sampling[J]. Journal of Parallel and Distributed Computing, 1992, 14(4): 361 - 372.

[12] PARK D, SINGH R, BAYM M, *et al*. IsoBase: a database of functionally related proteins across PPI networks[J]. Nucleic Acids Research, 2011, 39(Database issue): D295 - D300.