硕 士 学 位 论 文

基于蛋白质关系网络的蛋白质络合物抽取研究

Research of Protein Complex Extraction Based on Protein-Protein Interaction Network

作 者 姓 名：安波

学科、 专业：计算机应用技术

学　　号：20809335

指 导 教 师：王健副教授

完 成 日 期：2010-11-10

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

摘　　要

蛋白质络合物是由一组两个或多个蛋白质通过相互作用的关系而形成的蛋白质大分子结构，蛋白质络合物中的蛋白质共同完成一些特定的生物功能。蛋白质络合物在很多生物学过程中起到了关键的作用，是深入理解细胞组织和生物功能原理的基础。随着生物高通量技术的不断发展和广泛应用，大量的蛋白质相互关系被识别出来，这些数据为蛋白质络合物的抽取提供了新的视角。如何利用现有的蛋白质关系数据，借助于计算机工具抽取蛋白质络合物成为一个当前的研究热点。

本文首先介绍了蛋白质络合物抽取的相关知识和研究概况，然后介绍了蛋白质络合物抽取相关的评价方法,并针对现有研究出现的问题,提出了一个有效的蛋白质络合物抽取算法。

本文首先提出了一个基于关系网络修正的络合物抽取算法。主要利用关系网络的拓扑结构信息来递归计算关系网络中边的权值，根据得到边的权值信息对从蛋白质相互关系数据库中建立的关系网络进行去噪和完善。使用图聚类的算法对得到的修正关系网络进行聚类，对得到的完全子图中的节点的集合进行去重和融合，得到的结果是抽取到的蛋白质络合物。

其次，基于对蛋白质络合物生物特性的分析，为了合理利用蛋白质络合物的生物结构和功能特性，本文将蛋白质功能标注信息加入到算法中，用以计算蛋白质之间的功能相似度，利用这些相似度信息来处理蛋白质关系网络。并且根据蛋白质络合物的核心-附属关系特征，算法首先抽取到络合物的核心蛋白质集合。然后对这些核心蛋白质集合进行扩展，加入附属蛋白质，从而构成抽取的蛋白质络合物。

最后，本文针对蛋白质功能标注数据还不完善和抽取结果准确率比较低的情况，对算法进行优化。不仅仅在关系网络权值计算的时候使用标注信息，而且在对得到的完全子图进行过滤的时候加入标注信息。并引入机器学习的方法来对抽取到的蛋白质络合物进行过滤，选取合适的特征和分类器对抽取结果进行分类过滤，以提高最终结果的准确率。

总而言之，本文根据蛋白质关系网络的拓扑性质和生物性质，提出了一个从蛋白质关系网络中抽取蛋白质络合物的有效算法。本文的算法在多个蛋白质关系和蛋白质络合物数据集上进行了验证，取得了比现有方法更好的结果。

关键词：蛋白质相互关系；蛋白质关系网络；蛋白质络合物；蛋白质功能标注

Research of Protein Complex Extraction Based on Protein-Protein Interaction Network

Abstract

Protein complex is a macromolecular biological structure consisted of two or more associated proteins formed by interactions. And the proteins in the same protein complex perform biological function by cooperating with each other. Protein complex plays a critical role in many biological processes, so it is fundamental to the understanding of the principles of cellular organizations and biological functions. As a respond to wildly used of high throughtput approach involve genome-wide detection of protein interactions, the volume of protein-protein interactions is expanding at an incredible rate, which can help provide a brand new method to predict protein complexes. Therefore, it is a hot research area to utilize the protein-protein interactions to extract protein complexes.

In this paper, we first introduce the related knowledge and present research summary of extracting protein complexes. Then, we present the evaluation metrics of predicting protein complexes. And we propose an effective method to extract protein complexes from PIN.

First of all, an algorithm based on iterating revised protein interaction network is proposed. The method utilizes the topological characteristics to compute the weight of the edges in the network, and filtering the noisy interactions and adding high reliability interactions into the network. Then we extract all the maximal sub-graph from the revised network, filtering and merging the sub-graph with highly overlapped, the left sub-graphs are the predicting complexes.

In order to exploit the biological structure and functional properties of protein complexes, the algorithm incorporates the functional annotations into the detection of protein complexes. The function annotations are employed to compute the function similarity of proteins in the network, and processing the protein interaction network based on the similarity. According to the core-attachment structure of protein complexes, the method extracts the core proteins of the complexes and adds the attachment proteins into the clusters to form protein complexes.

Finally, this paper optimizes the algorithm according to the existing problems with the prediction of protein complexes from protein interaction network. Not only utilizing the protein functional annotations in the step of computing protein functional similarity, but also in the process of filtering non-qualified maximal sub-graphs. What's more, the machine learning method of SVM is introduced into this method. The algorithm selects proper feathers

and classifer to filter the extracted complexes to improve the precision of the prediction of protein complexes.

In conclusion, according to the topological characteristics of protein interaction network and the functional properties of protein complexes, we proposed a competitive algorithm to extract protein complexes from protein-protein interactions. We have applied our methods successfully on several protein-protein interaction databases and protein complexes databases.

Key Words：Protein-Protein Interaction; Protein Interaction Network; Protein Function Annotation

目　　录

1 绪论

1.1 研究背景

随着人类进入后基因时代，越来越多的基因、蛋白质及其相互关系被发现。如何分析和利用这些已知数据并使其为研究者服务成为当前的一个重要任务。蛋白质相互关系作为理解细胞功能、遗传、新陈代谢等生物过程的重要媒介，已经成为生物学家研究的一个热点。近些年，随着酵母菌双杂交技术[1]、质谱分析技术[2]、蛋白质芯片[3]等高通量技术的不断发展和广泛应用，生物学家们识别出了大量的蛋白质相互关系数据，生物专家和一些组织也建立了各种相应的蛋白质相互关系数据库。

传统的对蛋白质相互关系的研究中，生物学者是针对某个或者少量的蛋白质相互关系进行处理和分析[4]。由于相互关系被限制在一个比较小的范围里，因而分析得到的信息量比较少，这种方法在很大程度上依赖于生物学家对该领域的了解和生物实验。现在，生物学家通过各种实验得出结论，在生物体内发生的各种功能，例如：血细胞生成，有氧呼吸和细胞分裂等主要是以一些复杂的生物结构为单位的。例如，由一组蛋白质以及蛋白质之间的相互关系为单位的，这些蛋白质通过相互作用的关系协同完成一些特定的生物学功能。蛋白质络合物[5]是一种生物体内的重要大分子结构，是由一组两个或两个以上的蛋白质构成的具有特定功能的生物结构，通过蛋白质之间紧密的相互联系结合而成的，往往在功能上具有相似性，并且存在于相同的细胞位置。络合物中的蛋白质通过相互协作共同完成特定的生物功能。过去生物学家需要通过生物学实验的方法去抽取蛋白质络合物，需要进行大量的生物学实验，耗费大量的人力物力。因此，如何利用已知的蛋白质相互关系数据来建立蛋白质相互关系网络并从其中抽取蛋白质络合物，成为近几年生物网络分析的热点。

1.2 蛋白质络合物抽取算法的研究现状

目前，已经有多种抽取蛋白质络合物的方法，本文关注的主要方法是基于蛋白质关系网络的蛋白质络合物抽取算法。因而，在这里不讨论从生物学方法和其他的基于计算的蛋白质络合物抽取算法[6]。本章主要介绍近几年提出的比较有代表性的算法，以及能够取得比较好的抽取结果的算法。Dongen 等在 2000 年提出了基于随机游走的算法（MCL），该算法是基于图论的复杂网络聚类算法[7]。该算法首先根据得到的蛋白质相互关系数据构建邻接矩阵，在得到的关系网络中模拟流体流动，使用两个给定的操作来完成流动性区域的划分。算法会得到一些流动性比较大的区域，这些区域中的蛋白质集合就是抽取到的蛋白质络合物。算法第一步是根据蛋白质的相似性建立相似性矩阵，相

似度较大的节点之间被赋予较大的权重。第二步，使用扩展和膨胀算法来激发随机游走，这些都是通过矩阵变化来实现的。MCL 是一种通用性比较强的算法，广泛用于各种复杂网络的聚类研究，比如神经网络，社区网络等。Bader 等在 2003 年提出了基于局部密度的聚类算法（MCODE），该算法也是根据蛋白质相关系数据建立关系网络[8]。首先算法给每个网络中的节点根据网络中的局部密度计算并赋予一个权重，节点的权重随着节点所在的子图中的局部密度的增大而增大。在节点权重的基础上，使用贪心算法搜索网络中局部密度比较大的节点的集合，之后对得到的这些节点集合进行去重等后处理，就可以得到蛋白质络合物的集合。King 等在 2004 年提出了基于代价的邻接矩阵搜索聚类算法（RNSC）[9]。RNSC 首先定义了一个网络分割的评分方法，以对网络的分割做出代价评估。首先随机地分割网络图并继续重新分配结点， 直到获得最大的分割分数；然后将得到的节点的集合，根据它们的大小、绸密度和功能均一性进一步筛选，从而抽取出蛋白质络合物。RNSC 得到的聚类结果是一组互不相交的蛋白质络合物集合。Adamcsek 等在 2006 年提出基于图的拓扑结构信息的聚类算法（CFinder）[10]。这种算法主要依赖于蛋白质络合物在关系网络中是一种局部密度比较大的子图的特征。算法首先根据数据库中的蛋白质的相互关系构造关系网络；然后在这个图中使用贪心算法查找所有节点数大于等于 K（K 是预定义的值）的完全子图；然后合并具有共同的 K-1 阶完全子图的两个相邻的 K 阶完全子图，迭代的执行上述过程，直到没有满足合并条件的子图存在。这些合并后的子图，就构成了蛋白质络合物的集合。2007 年 Xiaoli Li 等提出了一个基于局部密度的算法（DECAFF）[11]，这个算法在查找局部密度比较大的子图的时候加入蛋白质功能标注信息，但是对蛋白质功能标注信息的应用主要是在最后的算法结果验证上，对得到的蛋白质络合物再进行合并然后再对得到的结果进行验证。Yanjun Qi 等在 2008 年提出了基于机器学习的蛋白质网络聚类算法（SCI-BN）[12]。这种算法提出了基于蛋白质络合物特征的方法，引入机器学习的方法对得到的蛋白质络合物进行分类，其中使用贝叶斯网络进行聚类。算法的主要过程是，首先使用一个启发式的算法对得到蛋白质关系网络中的密度比较大的子图，然后利用已知的标准的蛋白质络合物数据和一些随机生成的蛋白质集合对分类器进行训练，使用的分类器是基于贝叶斯原理的分类器。最后对上一步得到的子图使用贝叶斯分类器进行分类，这样就可以得到最终的结果，这个算法也取得了不错的结果。2008 年 Liu 等提出了基于最大完全子图的方法来进行蛋白质络合物的抽取（CMC）[13]，这个算法首先找到蛋白质关系网络中的所有密度比较大的子图，然后计算边的权重，根据权重对得到的蛋白质集合进行打分，根据这个分值对得到的蛋白质络合物进行排序，最后去掉那些重叠率比较高并且排名相对靠后的

蛋白质集合，剩下的蛋白质集合就是该算法在关系网络中抽取到的蛋白质络合物。2009年 Leung 等提出了基于蛋白质络合物的 Core-Attachment 结构的算法（CoreMethod）[14]，这个算法率先引入了由 Gavin 等提出的蛋白质络合物的 Core-Attachment 基本结构[15]，这个结构在 2008 年被 Luo 等人进行了验证，证实了蛋白质络合物确实存在这样的内部结构[16]。这个算法首先找到密度比较大的子图，让这些子图充当蛋白质络合物中的核心（Core）部分，然后通过定义一些规则加入附属（Attachment）蛋白质，这样就可以得到最终的蛋白质络合物。这个算法抽取的结果分为两个部分，一部分是蛋白质络合物，另外一部分是蛋白质络合物的核心集合。2010 年 Xiaoli Li 等提出了另外一个基于蛋白质络合物的 Core-Attachment 结构的蛋白质络合物预测算法（COACH）[17]，这个算法通过一系列复杂的数学计算，主要是计算蛋白质集合的 P 值，通过 P 值，找到蛋白质络合物的核心（Core）部分，然后也是通过规则的方法加入附属（Attachment）蛋白质。目前，这个算法在许多公开的蛋白质相互关系数据集上都取得了理想的效果。上述的几种算法都是在各种论文中被广泛引用的算法，有些算法，如 MCODE、CFinder、MCL 等算法是比较经典的蛋白质络合物抽取算法，并且具有比较广泛的通用性，另外一部分算法，如 SCI-BN、DECAFF、CMC、CoreMethod 和 COACH 属于近几年能够取得比较好的抽取效果的算法。这些算法都是在本文的试验中主要的对比对象。

1.3 待解决的问题

虽然已经提出许多基于蛋白质关系网络的蛋白质络合物抽取算法，并且，有些算法已经取得了不错的抽取结果，但是目前算法的抽取精度还不能满足生物学家的需要。本文认为，其中，一个比较重要的原因是现有的大多数算法，都仅仅是基于蛋白质络合物内部的蛋白质连接紧密的特点，而忽略了蛋白质络合物的一个重要的生物特性，也就是蛋白质络合物中的蛋白质具有相同或相似的功能。例如，现在有许多公开的蛋白质标注数据集，而上述的算法很少或者根本没有利用这些蛋白质功能标注数据库。当然，除了蛋白质功能标注之外，从蛋白质关系网络中抽取蛋白质络合物还面临着许多难点。例如，对结果的验证，现在主要结果验证方法是与已知的蛋白质络合物数据库进行对比，但是由于现有的蛋白质络合物数据库也是不完整的，从而使得实验结果很难比较真实的展示出来，另外如何跟生物学家合作，使得抽取的结果能够通过生物实验做进一步的验证，将是一个急需解决的问题。下面列出了一些其他在试验中遇到的问题：

（1） 先有的蛋白质相互关系数据集中的关系置信度不高。高通量的蛋白质关系预测技术的广泛应用产生了大量的错误数据。例如，据[18]称，基于高通量的酵母菌双杂交

系统产生的数据仅有50%左右是可信的。使用这些蛋白质相互关系数据建立的蛋白质关系网络，必然会有很多的噪音数据存在，从而使得蛋白质络合物的抽取结果也会下降。

（2） 一个蛋白质可以有多个不同的功能。因为一个蛋白质可能具有不同的功能，所以一个蛋白质很有可能会出现在多个不同的蛋白质络合物中，这就使得早期将蛋白质络合物定义为不相交的蛋白质集合的方法无法得到理想的效果。

（3） 功能不同的蛋白质之间也会存在相互关系。这种蛋白质之间随机性的相互关系，加大了蛋白质关系网络的拓扑关系复杂性，增加了蛋白质络合物预测的难度。通过对蛋白质络合物的分析可知，蛋白质络合物中的有些蛋白质与其他的蛋白质络合物没有相同的功能标注，但是经过生物实验证明也是蛋白质络合物的一部分。

（4） 蛋白质功能标注信息不断增多，如何利用已知到蛋白质功能标注信息来改善蛋白质络合物抽取的结果，也是一个急需解决的问题。

1.4 本文工作

本文主要阐述的是基于网络模型的蛋白质络合物抽取技术。针对现有的基于蛋白质关系网络的蛋白质络合物抽取中出现的问题，比如：蛋白质关系网络数据高错误率、未能很好的利用蛋白质络合物本身的结构特征和生物特性，提出了相应的解决方法。

论文中使用的网络模型有助于我们将蛋白质相互关系从复杂的生物功能提取出来，并建立一个数学模型。使得蛋白质络合物的抽取转换成对复杂网络的聚类。这样做的优点有两方面：一方面从复杂的生物分析中解放出来，转换为可以用计算机处理的数学模型；另一方面，增强了算法的通用性，通过建立关系网络的方法来进行蛋白质络合物聚类和分析的算法，可以用到很多具有类似特点的数据挖掘中，例如 Pathway 的预测，社会关系网络中的社区发现等。

针对现有研究中，出现的主要问题，设计合理的算法对蛋白质络合物抽取的整个过程进行优化。首先对蛋白质关系数据进行处理，实现对关系网络的去噪。然后对关系加入蛋白质络合物的核心-附属内部结构，使用蛋白质功能标注信息对网络中的节点进行分类处理。最后借助于机器学习的方法，对得到的蛋白质络合物进行过滤，一步步的来解决现存的各种问题。

1.5 本文的结构

论文共分为五章，详细阐述了基于蛋白质关系网络的蛋白质络合物抽取算法，具体的问题分析，实验设计以及性能评估，具体章节安排如下：

第一章，绪论，综述了论文所研究内容的背景以及研究现状，介绍了本文研究领域中的主要方法和论文的结构安排。

第二章，主要介绍本文使用中使用的蛋白质相互作用关系数据库和蛋白质络合物抽取的相关知识以及实验的数据集，最后介绍常用的蛋白质络合物抽取算法的评价指标。

第三章，详细介绍基于蛋白质相互关系迭代计算和关系网络预处理的抽取方法，以及在试验语料上的结果及其分析。

第四章，详细介绍使用蛋白质功能标注对蛋白质关系数据的过滤和置信度计算，以及加入蛋白质络合物中的 core-attachment 结构给蛋白质络合物抽取带来的影响。

第五章，详细介绍一种优化蛋白质功能标注利用率的算法来改善蛋白质络合物的抽取效果，并加入了机器学习的方法对得到的蛋白质络合物进行过滤。

论文的总结，介绍了本文研究的主要内容、工作及下一步的工作。

2 相关知识及评测指标

2.1 蛋白质相互关系及关系网络模型

由上一章可知，目前已经有很多的蛋白质相互关系数据公开并提供给研究者免费下载和使用。蛋白质及其相互作用关系在大多数生物学过程中起到至关重要的作用。一般来讲，蛋白质很少单独的完成所拥有的功能，更多的是通过与其他蛋白质相互协作来共同完成一些生物功能。因此，蛋白质相互关系是理解生物学进程的关键因素。

为了能够大规模的分析蛋白质相互关系，生物专家们提出了一些生物关系网络。其中，蛋白质相互关系数据可以表示成一个网络模型 G =（V，E）；其中 G 代表整个蛋白质关系网络；V 表示数据库中的蛋白质集合，每个节点代表一个蛋白质；E 表示数据库中的蛋白质之间的相互关系，每个边代表一对蛋白质之间的相互关系。这样对于蛋白质络合物的预测就可以转换为蛋白质关系网络中的聚类问题，也就是从给定的蛋白质关系网络中抽取一些满足特定要求蛋白质的集合。

一般情况下得到的关系网络是一个无向的无权图，因为现有的数据库很少会给出蛋白质相互关系作用的方向和置信度。

2.2 蛋白质关系网络的基本性质

（1） 节点的度：在一个无向图中，节点的度是与这个节点直接相连的节点的个数。在有向图中，每个节点有两种度—入度和出度。节点的入度是所有直接指向这个节点的节点个数；节点的出度为这个节点直接指向其他节点的边的个数。

（2） 节点间的路径：两个节点之间的路径 Path(u，v)被定义为：在关系网络中从节点 u 出发，经过一系列的边和节点后到达节点 v。路径的长度为在这个路径上边数的个数。最短路径是两个节点所有路径长度的最小值对应的路径。

（3） 小世界性[19]：蛋白质关系网络因为节点和边个数比较多，结构非常复杂。但是它本身也有着一些特殊的性质，例如在图中所有节点之间的最短路径的平均值是很小的。也就是说在关系网络中大多数节点之间的最短路径长度都比较小。这个性质是被 Watts 和 Strogatz[20]验证过的。

（4） 无尺度分布性：在关系网络中节点度的分布，也就是给定的一个节点的度为 k 的概率呈一个幂律分布。这个性质表明，关系网络中只有很少一部分节点拥有较大的度，这些节点也成为一种特殊的节点—中心节点（hub node）。

2.3 关系网络的价值

生物专家根据蛋白质关系网络和蛋白质络合物的特征，给出了一些对蛋白质关系网络研究特别有用的结构特征。

（1） 在关系网络中有相互关系的蛋白质具有相同或者相似的功能。因此，我们可以通过分析与某个未知功能的蛋白质相连的已知功能的蛋白质或者络合物来判断这个蛋白质的功能。在聚类的过程中也会比较关注具有相同功能标注的蛋白质。

（2） 在关系网络中，局部密度比较大的子网很有可能会构成一个蛋白质络合物。现有的几乎所有的算法都会通过找到一个密度比较大的子图作为后续处理的集合，因此这个性质也为蛋白质络合物的抽取提供了一个必要的前提。

（3） 分析关系网络的拓扑特性可以更好的理解生物系统，例如 Pathway 的一些结构特征等。

2.4 蛋白质络合物的结构特性

蛋白质络合物是由在相同时间和位置发生相互关系的蛋白质组成的能够完成一些特定功能的大分子结构。位于一个蛋白质络合物中的不同蛋白质不仅具有相同或相似的功能，根据 Gavin[15]在 2006 年经过大量验证后提出的理论可知，蛋白质络合物本身具有一些结构特征，蛋白质也被划分为三种，分别是核心（core）、组块（module）和附属（attachment），如图 2.1 所示。这三类蛋白质的特征如下：

（1） Core-protein：这部分蛋白质集合之间的功能非常相似，并且它们之间的相互联系也很大，反应在关系网络中就是子图的密度比较大。

（2） Module-protein：这部分蛋白质集合的特点是，这些蛋白质会一起出现在一个或者多个蛋白质络合物中。但是实验证明这些蛋白质一般和 core-proteins 一起出现，只出现在一个络合物中。

（3） Attachment-protein：这类蛋白质可以出现在一个或者多个不同的蛋白质络合物中，而没有其他的约束。

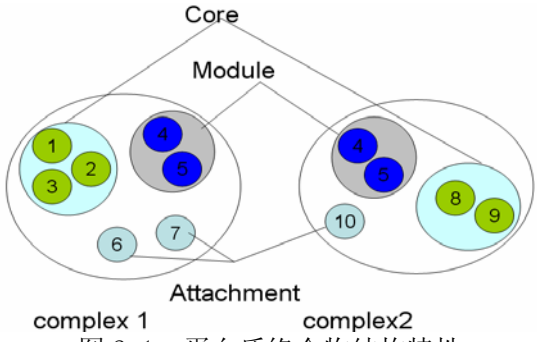

图 2.1 蛋白质络合物结构特性

Fig. 2.1 The structure of protein complex

2.5 实验语料

2.5.1 蛋白质相互关系数据库

为了充分验证算法的结果，本文使用了多种公开的蛋白质关系数据集。但是因为这个领域还没有公开的评测，所以并没有标准的数据集提供本文使用。论文中主要选取了在各种论文中多次出现的蛋白质关系数据集，以方便对其他算法验证和进行对比试验。下面列出本文使用过的蛋白质相关关系数据集。

（1） BioGrid 蛋白质相互关系数据库[21]：该语料来自 Biological General Repository for Interaction Datasets 数据集，是 2003 年由一些生物专家手工搜集的包含蛋白质和基因之间相互关系的跨物种的数据集，因此该数据集具有很强的置信度。不过，在后续的一些版本的数据集中已经加入了来自高通量方法的数据。

（2） Gavin06 蛋白质相互关系数据库[15]：该语料来自 Gavin 等专家在 2006 年提取出来的蛋白质相互关系，这个数据集也具有高度的置信度，并且在很多的方法中被使用。

（3） DIP 蛋白质相互关系数据库[22]：该语料来自 Database of Interacting Proteins，这个语料是由加州大学洛杉矶分校发起收集的，现在已经成为蛋白质相互关系数据集中很有权威的数据集，并在大量文献中引用。

（4） Krogan 蛋白质相互关系数据集[23]：该语料是由 Krogan 等专家负责搜集并发布的，在很多的算法中被引用。

这四种蛋白质相互关系数据集的基本信息在表 2.1 中列举出来，包括数据集中蛋白质的个数、蛋白质相互关系的个数和每个蛋白质的平均度。

表 2.1 蛋白质相互关系数据库详细信息

Tab. 2.1 The detail information of protein interaction databases

数据库名称	蛋白质个数	蛋白质相互关系个数	蛋白质的平均度
BioGrid[21]	1376	10918	15.87
Gavin06[15]	1430	6531	10.62
DIP[22]	4930	17203	6.98
Krogan[23]	3581	14077	7.86