# 绪论

## 背景部分

随着人类基因组计划的完成，人类对于生命科学的理解进一步加深，并由此衍生出了大量的研究方向，而系统生物学作为其中的一门新兴学科，获得了研究者的广泛关注。跟传统的生物学相比，系统生物学不仅仅只局限于单个蛋白质，单个基因对于生物机体产生的影响，而是通过研究蛋白质和蛋白质，基因和基因，蛋白质和基因，蛋白质和疾病之间的相互作用和相互关系，利用计算机和数学建模，通过计算的方式来预测细胞，器官甚至整个生物机体的系统表现。可以看出，生物实体之间的基本关系作为系统生物学最重要的基础部分，对于更深层次去理解机体的表现具有重要的意义。

21世纪以来，随着生物技术的迅猛发展，相关的生物文献也以一种爆炸式的方式在增长，这些生物文献大多是以电子版的方式存储在生物文献的数据库中，全世界的科研人员可以通过检索这些数据库，从而获得最新的研究成果。以最出名的Medline数据库为例子，在过去的十年时间里，它收录的文档总数翻了一倍，目前该数据库已经涵盖了70多个国家地区的7000多种期刊，收录接近2300万的生物文献，基本上涵盖了整个生物临床医学和生命科学的范围。值得注意的是，这些文献大多是以纯文本的，非结构化的方式存储在数据库中，这些海量的数据加上非结构化的存储方式，使得高效的检索成为技术难题。所以如何利用自然语言处理和数据挖掘的方式去分析这些海量文本，从原始文本中提取有用的信息成为了一个研究热点。



图1：基于关键词“breast cancer”的PubMed的检索结果

PubMed作为最大的，免费的生物文献搜索引擎，基本上涵盖了Medline，PMC，OLDMEDLINE，record in process等几个主流的生物文献的数据库，进一步的，它能提供各种生物医学论文的搜索以及摘要展示，可以为生物领域的研究者们提供最前沿的研究动态。通过收集用户的搜索记录，PubMed能够提供比较友好的检索体验。以关键检索词“breast cancer”为例，图1展示了PubMed的检索结果。从图中可以看出，在键入关键词“breast cancer”以后，搜索引擎能够联想出很多关于breast cancer的主题，比如breast cancer的治愈，breast cancer的风险等等。但是更深层次的主题，比如想要获取breast cancer相关的基因，如果没有丰富的生物医学相关的知识，很难直接从搜索引擎获得这些基因的信息。而生物实体关系抽取技术的研究，为提高搜索引擎的搜索精确性提供了一个很好的技术支持，如果关系抽取系统已经抽取了所有与breast cancer相关的实体，那么搜索引擎技术就可以从这些实体中选择合适的内容，为用户提供更加精确的搜索结果，因此，高质量的，精确的，自动的生物关系抽取系统是有必要的。

## 生物实体的关系抽取