**生物信息学**（ **Bioinformatics**）是在生命科学的研究中，以[计算机](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=430164)为工具对[生物信息](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=452396)进行储存、检索和[分析](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=135054)的[科学](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=35999)。它是当今生命科学和[自然科学](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=46523)的重大[前沿领域](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=70250723)之一，同时也将是21世纪自然科学的核心领域之一。其研究重点主要体现在[基因组学](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=250454)（Genomics）和[蛋白质组学](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=161481" \t "_blank)（Proteomics）两方面，具体说就是从[核酸](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=261113)和[蛋白质](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=78893)序列出发，分析序列中表达的结构功能的生物信息。

生物信息学的发展共经历了三个时期，分别是：前基因时代，基因时代和后基因时代。前[基因组](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=142526)时代（[20世纪90年代](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=530420)前） 这一阶段主要是各种序列比较算法的建立、[生物](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=105252)[数据库](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=60685)的建立、[检索工具](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=7846097)的开发以及DNA和[蛋白质](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=78893" \t "_blank)序列[分析](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=135054)等。基因组时代（20世纪90年代后至2001年） 这一阶段主要是大规模的基因组测序，[基因识别](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=10923053)和发现，[网络数据库](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=450370)系统地建立和[交互界面](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=72483844)工具的开发等。后基因组时代（2001至今） 随着[人类基因组](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=161566)测序工作的完成，各种[模式生物](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=541808)基因组测序的完成，[生物科学](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=82123692)的发展已经进入了后基因组时代，[基因组学](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=250454)研究的重心由基因组的结构向[基因](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=517296)的功能转移。这种转移的一个重要标志是产生了[功能基因组学](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=822030)，而基因组学的前期工作相应地被称为[结构基因组学](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=7840264)。

本组对于生物信息学的研究主要应用于从蛋白质的序列信息出发，对蛋白质序列上的蛋白质相互作用位点进行预测。蛋白质是细胞功能的最终执行者，它对于维持正常的生命活动发挥着至关重要的作用，随着人类基因组计划的全面推进和越来越多的模式生物基因组全序列测定的完成，后基因组时代已经到来。后基因组时代的主要目标之一就是要阐明细胞体系中生物大分子之间的相互作用机制。而蛋白质相互作用是所有生命活动发生的基础，是细胞进行一切代谢活动的必要条件。对蛋白质相互作用进行研究不仅有助于解释生命活动的本质，而且对疾病发生机制的了解及有效药物的开发均起到推动性作用。因此，蛋白质相互作用研究必将成为生命科学领域的又一个研究热点。

关于蛋白质相互作用的研究方法大体可分为两类，即实验方法和生物信息学方法。由于人类基因组计划的实施使得蛋白质序列数据库中的数据急剧上升，使得利用实验的方法一一测定蛋白质的结构以及它们之间的相互作用所花费的成本太高，实验周期太长。因而有必要发展一些可靠的理论预测方法，借助于计算手段来得到某种程度的解决。也就是从蛋白质序列信息出发，预测蛋白质相互作用和功能：分析预测基因在翻译表达方面的调控机制；以及建立细胞水平的新陈代谢模型等等。因此用于蛋白质-蛋白质相互作用计算分析方法对于实验方法的辅助作用将变得越来越重要。

计算分析方法主要有两个方向：宏观水平上，集中研究蛋白质相互作用网络。微观水平上，着重了解相互作用机制和相互作用位点的预测。随着基因组数据的增长，对蛋白质-蛋白质相互作用网络的研究已经取得飞快进展，这些研究为我们描绘出了大量的蛋白质-蛋白质相互作用网络，而对于微观水平即从蛋白质序列或蛋白质结构出发预测相互作用位点仅有少量研究。#而无数个残基-残基相互接触会促进蛋白质-蛋白质相互作用，一旦识别出产生蛋白质-蛋白质相互作用的残基，我们将很容易知道这些残基与哪些蛋白质发生相互作用。识别相互作用位点是译解蛋白质功能机制的一把钥匙，对于全面理解分子过程以及药物设计、药物发展起着决定性作用。#

20世纪90年代以来，国内外研究者们提出了大量的计算方法用于预测蛋白质相互作用及其位点，其中主要有：线性回归(linear regression)、得分函数法(scoring function)、支持向量机(support vector machine,SVM)、神经网络(neural network)、朴素贝叶斯(Naive Bayesian)及贝叶斯网络、隐马尔科夫模型(Hidden Markov model,HMM)等。

在对相互作用界面进行属性分析或对位于界面上的残基即相互作用位点进行预测之前，需要依据蛋白质复合物的结构信息并运用计算方法对这些关键位点进行定义。目前，相互作用位点的定义方法存在多种，并没有统一的形式。而相互作用位点的定义方式也直接关系到属性分析的结果和预测方法的效果。大多数学者采用的策略是先将一个蛋白质包含的残基分为两类，一类是表面残基，另一类为被包埋的残基。其中，只有表面残基才有可能与其他蛋白质所包含的残基发生物理上的接触。要得到表面残基则需要计算各个残基的相对溶剂可及表面积。计算这一指标的软件主要包括两种，分别为NACCESS和DSSP。当残基的相对溶剂可及表面积大于一个阈值时，则定义它为一个表面残基。在得到表面残基后，通过计算它们和与其相互作用蛋白质间的院子距离来定义界面残基。除了上述这种定义相互作用位点的策略之外，目前使用较广的还有另外两种方式，这两种方式则不定义表面残基，而是直接把蛋白质包含的残基分为界面残基和非界面残基。其中，一种是通过计算蛋白质在结合前和溶剂可及表面积的改变量来定义，另一种则是通过计算几何学中的Voronoi图方法来定义。虽然这些定义方式各不相同，但已有研究表明通过不同方式得到的相互作用位点，其中大部分是相互重叠的。

截止到2006年，在蛋白质数据库（Protein Data Bank,PDB）中收录有39204种生物大分子的结构，这其中有超过35000个记录是有关蛋白质的三维结构的，这些蛋白质主要是通过X射线衍射（X-ray）、核磁共振（NMR）和电子显微镜（cryo-EM）等多个实验方法以不同的分辨率所测定的。但与此同时，在蛋白质相互作用有关的数据库（Database of Interacting Proteins，DIP）中却只有几千条蛋白质相互作用的记录。虽然现在针对蛋白质相互作用的实验方法在速度和精度上都在提高，但相对于蛋白质的结构测定速度来说，还是远远不够的。

由于完整的蛋白质相互作用组是一个动态的研究对象，除了不同细胞之间的差异，一个特定细胞内的蛋白质相互作用组也是随时间变化的。虽然人们知道生命活动的进行是依赖于细胞内蛋白质之间的相互作用，但人们如何来确定各种各样的蛋白质之间是否存在相互作用，如果存在又在什么条件下发生，仍然知之甚少。这种情况只是在近几年才有所改观，由于高通量生物技术的快速发展，才使人们得到大数据量的蛋白质数据和蛋白质相互作用数据，从而使人们从系统的高度来研究蛋白质相互作用成为可能。

在前基因组时代，免疫共沉淀（Co-Immunoprecipitation）是测定蛋白质-蛋白质相互作用的最基本方法；在后基因时代，我们需要发展分析蛋白质相互作用的高通量方法。目前生物技术的高通量方法主要有：双向凝胶电泳、酵母双杂交技术、微阵列和蛋白质芯片技术和质谱分析技术等等。这些先进的、大规模分析蛋白质相互作用技术的发展推动了蛋白质组学的进步。但是这些技术都受到当前实验科学的制约，耗时耗力，又受到蛋白质相互作用的时间和空间制约，而且高通量技术在检测出大量蛋白质相互作用的同时，也带来了很多“假阳性”和“假阴性”数据的困扰。

与此同时，为了解决生物实验技术的缺陷，辅助生物实验的进行，并对由实验所获得的相互作用数据进行分析，使用计算方法来研究蛋白质相互作用就成了当前计算机科学、数学、生物学等相关领域的研究人员所研究的一个热点。近几年来，有很多的计算方法被发展出来预测蛋白质之间的相互作用或对蛋白质发生相互作用的位点进行定位。这些方法从不同的生物学背景知识出发，如基因组信息、蛋白质家族进化相关信息、蛋白质的一级结构或三维空间结构信息等等，来对蛋白质的相互作用进行了多方面的研究。这些方法的使用为研究蛋白质相互作用提供了一条有效、简洁和快速地建立细胞内蛋白质相互作用网络的途径。

运用计算方法研究蛋白质相互作用的研究过程可分为两个阶段。第一个阶段的研究侧重于相互作用界面的属性分析，通过收集不同类型蛋白质复合物的结构信息，对各类相互作用界面的属性进行对比分析，从而得到了一些具有普遍性的界面特征。第二个阶段则更多地关注相互作用位点预测方法的开发，随着统计学习理论的逐步完善，利用机器学习算法并结合已知的各种属性特征对相互作用位点进行识别已成为当前的研究热点。

在这个过程中，做出突出贡献的研究主要有：Kini和Evans观察到蛋白质-蛋白质相互作用位点附近脯氨酸残基总是频繁出现，于是通过检验“脯氨酸支架（proline brackets)”的存在对潜在的蛋白质-蛋白质相互作用位点进行了预测；Jones和Thornton基于六种特征参数，对蛋白质-蛋白质相互作用位点进行了系统分析，采用一种记分功能，成功地预测了具有59种结构的蛋白质数据集的接触面。Gallet等人基于Eisenberg等人提出的膜和蛋白质表面片段检测方法，通过分析蛋白质序列的疏水性进而识别出蛋白质-蛋白质相互作用位点。还有其他一些依靠多序列比对检测保守残基或者关联突变的方法预测蛋白质-蛋白质相互作用位点，尤其近几年来，对蛋白质-蛋白质相互作用位点的研究已经取得相当大的进展，方法和手段也越来越趋向多元化和信息化，Yan等人基于支持向量机和贝叶斯等分类算法，利用相邻残基序列轮廓（sequence profiles）信息成功地预测了蛋白质-蛋白质相互作用位点上的接触面残基；Asako Koike和Toshihisa Takagi基于支持向量机分类算法，将空间结构相邻残基序列轮廓（sequence prorile)信息和序列相邻残基序列轮廓（sequence profiles)信息应用于蛋白质-蛋白质相互作用位点的预测，取得了较好的结果。