常用计算机辅助药物设计软件教程

主编　张亮仁

CHANGYONG JISUANJI FUZHU YAOWU SHEJI RUANJIAN JIAOCHENG

中国医药科技出版社

常用计算机辅助药物设计软件教程

主　编　张亮仁
副主编　刘振明
编　委　张　双　吴星宇　王玉飞　赵　亮
沈燕君　罗棋耀　夏　杰　裴　芬
陈　亚　曾凌晓　薛喜文

中国医药科技出版社

内 容 提 要

计算机辅助药物设计已成为一门新兴的研究领域，在合理药物设计中发挥不可或缺的作用。本书依据实际工作中各种软件的应用，介绍了分子动力学模拟、分子对接、蛋白质同源模建、定量构效关系的研究、药效团的构建、虚拟筛选等内容，且每一内容有不止一个软件的使用介绍，适合入门计算机辅助药物设计的广大师生参考使用。

图书在版编目（CIP）数据

常用计算机辅助药物设计软件教程 / 张亮仁主编. —北京：中国医药科技出版社，2017.4

ISBN 978-7-5067-9215-8

Ⅰ.①常… Ⅱ.①张… Ⅲ.①药物-计算机辅助设计-教材 Ⅳ.①R914.2-39

中国版本图书馆 CIP 数据核字（2017）第 067021 号

美术编辑 陈君杞

版式设计 张 璐

出版 中国医药科技出版社

地址 北京市海淀区文慧园北路甲 22 号

邮编 100082

电话 发行：010-62227427 邮购：010-62236938

网址 www.cmstp.com

规格 710×1000mm 1/16

印张 16½

字数 257 千字

版次 2017 年 4 月第 1 版

印次 2017 年 4 月第 1 次印刷

印刷 三河市双峰印刷装订有限公司

经销 全国各地新华书店

书号 ISBN 978-7-5067-9215-8

定价 42.00 元

前　言

从20世纪90年代开始，随着计算机技术的迅速发展以及药物化学、分子生物学和计算化学的发展，计算机辅助药物设计（CADD）也快速发展起来，成为一门新兴的研究领域。与此同时，CADD的发展和应用，也大大促进了药物设计和新药开发的效率，CADD现在已经成为合理药物设计中不可或缺的一环，在药物设计中起着越来越重要的作用。因此，CADD方法的理论和应用研究具有非常重要的意义。关于CADD的理论，很多教科书和专著已进行大量介绍，适合初学者学习理论知识，但是在实际中的操作，却鲜有基础书目可以参考。本书编者长期从事相关工作，根据在实际工作中对各种软件的应用，编写了本书，介绍计算机辅助药物设计时需要用到的部分软件的使用，以实现设计目的为导向，着重介绍软件的操作方法，使用过程中的细节与原理，可以使初学者迅速入门。本书内容包含了分子动力学模拟、分子对接、蛋白质同源模建、定量构效关系的研究、药效团的构建、虚拟筛选等内容，且每一内容有不止一个软件的使用介绍，希望能对入门计算机辅助药物设计的广大师生有所帮助。

本书的编写是从初学者的角度，手把手、一步一步介绍各软件的使用方法、功能，并配有图片，使初学者能够顺利学习、使用各种常用软件。

限于编者的水平有限，书中可能存在不妥之处，敬请广大读者批评和指正。

编　者

2017年1月

目　录

第一章

同源模建法预测蛋白质结构

第一节　同源模建概述

蛋白质三维结构很大程度上决定了蛋白质的功能，因此获得蛋白质的结构并对其进行分析是现代分子生物学的重要课题。从药物分子设计的角度考虑，大多数药物靶标都是蛋白质，因此得到足够精确的蛋白质结构对于药物分子和靶点之间的相互作用研究及基于结构的药物设计都是非常重要的。目前，通过实验方法如 X 射线晶体衍射法和 NMR 法已经测出大量的蛋白质及其复合物的结构，但与已测得的蛋白质序列相比还是有很大差距，这也大大影响了人们对蛋白质结构和功能关系的研究。多年来研究者一直试图从序列信息来预测蛋白质的三维结构，目前主要采用的是同源蛋白预测的方法，即同源模建（homology modeling）。

1969 年，Browne 及其同事首先以鸡蛋清溶菌酶的结构为基础，手工模建了 α-乳白蛋白的空间结构，并获得了成功，从此开创了利用同源模建技术预测蛋白质空间结构的先河。1981 年，Greer 等人建立了利用多个同源蛋白质进行结构预测的方法，并将该方法用于模建哺乳动物丝氨酸蛋白酶的结构。

同源模建，也称为比较模建（comparative modeling），是蛋白预测技术中最重要的一门技术。其基本假设是序列的同源性决定了三维结构的同源性。一个未知结构的蛋白质分子（目标蛋白）的结构可以通过与之序列同源且结构已知的蛋白（模板蛋白）来进行预测。一般情况下，如果目标蛋白序列与模板蛋白序列之间的同源性在 50%以上时，那么通过模板蛋白搭建出来的蛋白结构具有很高的准确性；如果序列之间的同源性在 30%～50%时，通过模板蛋白搭建出来的蛋白结构准确性较高；如果目标蛋白序列与模板蛋白序列同源性在 30%以下时，所得目标蛋白结构的可信度较差，很难得到较好的结果。

同源模建的基本步骤主要包括同源蛋白的搜索和模板的选择、序列比对、模

型的建立、模型优化与模型评价五个部分。

（1）同源蛋白的搜索和模板的选择：从已知三维结构的蛋白质数据库（PDB）中搜索、挑选与目标蛋白序列相似的结构作为模板，一个目标蛋白的不同结构部分可以采用不同的模板模建。

（2）序列比对及确定结构的保守区：序列比对（sequence alignment）是同源模建的关键部分，也是最复杂和最困难的部分。序列比对分为多重序列比对和结构比对。多重序列比对可以确定序列相似性片段；结构比对可以确定结构上的保守片段，如跨膜螺旋的位置等结构特征。结合有关实验结果可以调整序列比对的结果。如果目标蛋白有两个以上已知结构的参考蛋白，可通过这些参考蛋白之间的结构叠合来确定结构保守区；如果参考蛋白中只有一个是有空间结构的参考蛋白，那么结构保守区的确定就必须通过多重序列比对的方法来实现。

（3）模型的建立：一般分为蛋白主链的模建和侧链的安装两步。结构保守区的主链坐标可以直接由参考蛋白拷贝下来，主链的模建主要在于环区。一般环区的模建有两个途径，一是片段搜索，二是自动生成法。侧链的安装主要是通过搜索构象库挑选出最佳的侧链构象组合。建立目标蛋白分子骨架的三维空间坐标，不同的同源模建方法的区别就在这里。目前用的最多的是通过同源蛋白间的序列比对来确定骨架结构，此外还有片段匹配法和几何限制法等。

（4）模型优化：蛋白质分子的主链和侧链都确定后只能说得到了一个初步结构，需要进一步优化。优化的目的是用来消除原子间的重叠以及不合理的构象，尤其是柔性区的构象。优化一般采用分子力学与分子动力学的方法。

（5）模型评价：用于模型评价的指标有很多，主要包括立体化学和能量评价两个方面。通过评价可以检测出模建的蛋白中哪些残基的构象不合理，不合理的程度有多大，这样研究者就可以依此对不合理的部分进行再优化或改变策略重新模建。

目前，同源模建方法构建的三维结构，误差主要有氨基酸侧链的空间位置，比对不完全正确部分的结构扭曲和没有序列比对部分的较长的 loop 区。尽管同源模建方法有明显需要改进和提高的地方，但在解决实际问题上发挥了实际的功效。随着新发现的蛋白质一级序列和蛋白质折叠方式在数量上的快速增长，同源模建方法显得越来越重要。

目前，用于同源模建的在线服务器和软件有很多。在线服务器如 SWISS-MODEL，EsyPred3D，CPHmodels 等，灵活性低，按照固定的算法难以得到符合实际的模型。商业软件包括如 Sybyl 和 Discovery Studio 等中都有用于同源模建的

相关模块，其中 DS 中的同源模建主要是基于 Modeler 程序。Modeler 是目前使用最为广泛，预测最为准确的同源模建工具之一。

第二节　同源模建

Accelrys 公司的同源模建（Discovery Studio，简称 DS）是基于 Windows/Linux 系统、面向生命科学领域的分子建模和模拟环境。Discovery Studio 针对生命科学应用，提供生物大分子及有机小分子建模的显示工具、功能分析工具、结构改造工具、动力学模拟工具等，帮助研究人员在实验前全面了解生物分子的结构与功能，从而有针对性地设计实验方案，提高实验效率，降低科研成本。

Discovery Studio 为用户提供了一整套利用 Homology Modeling 方法自动预测蛋白质空间结构的工具。用户只需要提供蛋白质的氨基酸序列就可以轻松完成模型构建及模型可信度评估的工作。DS 的同源模建主要基于 MODELER 程序。以下实例为一个人肾上腺素受体 β 亚型的同源模建过程，选用 Discovery Studio 2.5 进行同源模建，内置 Modeler version 9v4。

一、搜索并识别模板

先在 uniprot(http://www.uniprot.org/)或者 pubmed（http://www.ncbi.nlm.nih.gov/protein/）中找到目标序列，使用 BLAST 在 Protein Data Bank（PDB）数据库中搜索模板。

1. 载入序列

打开 DS2.5，选择 File | New | Protein Sequence Window，把目标序列粘贴到窗口中，或直接在 DS2.5 中打开目标序列的 fasta 文件（图 1-1），注意可先更改简化序列命名，默认的有基因名称可能会使后续步骤不能成功进行。

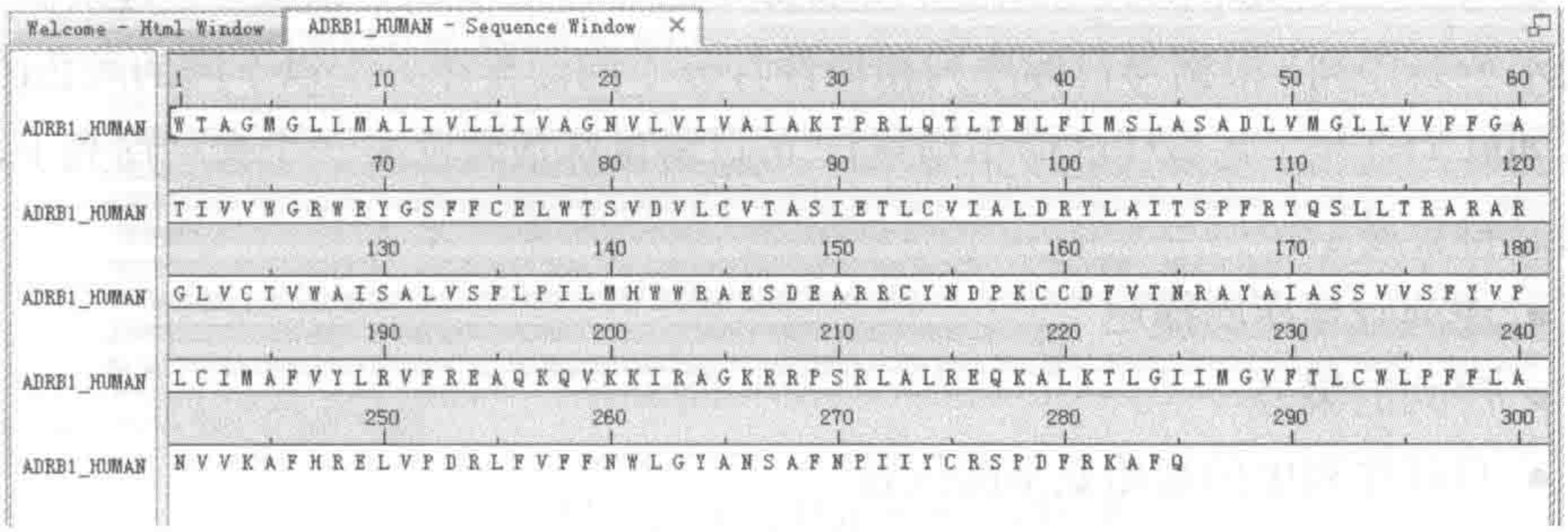

图 1-1　读入的序列

2. BLAST search

在 protocols 中，展开 Sequence Analysis，双击 BLAST Search（NCBI Server），出现参数选择的对话框。在 Input Sequence 中选择目标序列，在 Input Database 中选择 pdbaa（图 1-2）。

BLAST Search (NCBI Server) ×

Parameter Name	Parameter Value
Input Sequence	ADRB1_HUMAN:ADRB1_HUMAN
▷ Input Database	pdbaa
Scoring Matrix	BLOSUM62
Gap Penalties	Existence: 11 Extension: 1
Gapped Alignment	True
E-value Cutoff	10
Use Composition Based Statistics	True
Filter Low Complexity	True
Maximum Hits	250
▷ Advanced	

图 1-2　BLAST Search 参数设置

点击▶或按 F5 键运行 protocol。完成后，会显示 Job Completed 对话框（图 1-3），点击 OK 将其关闭。

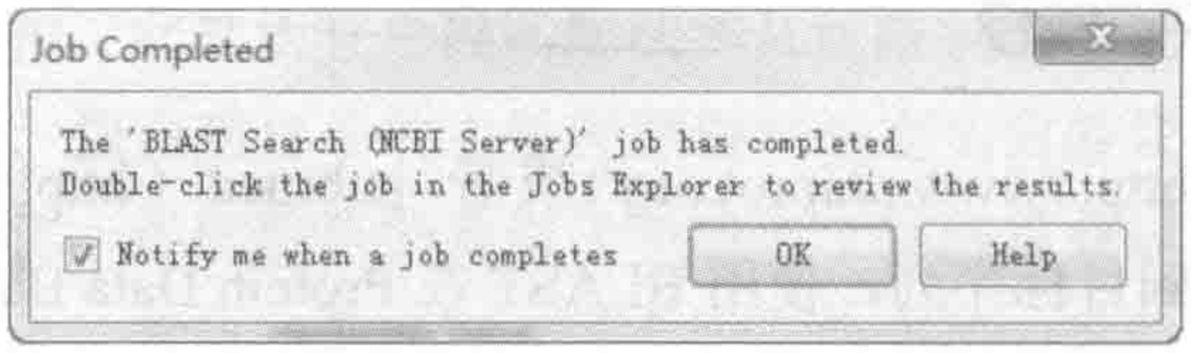

图 1-3　Job Completed 的对话框

查看结果：在 Jobs 中，双击完成的 protocol，打开一个 Html 窗口，里面包含 Report.htm 文件（图 1-4）。在 Output Files 部分，点击 ADRB1_HUMAN.xml，将打开 BLAST 搜索的结果。

默认打开为 Map View（图 1-5），将命中结果都显示在一张图中，每条横条线框表示一条序列，根据与目标序列相似性打分不同而排序并配以不同的颜色(分数超过 400 为红色，是最佳的命中结果)。可以将鼠标放置在某一个命中序列上，如下信息将会显示：

- 序列数据库的描述
- 序列的编号
- 目标序列中的起始氨基酸位置
- 数据库中命中序列的起始氨基酸位置

- 命中序列的长度
- 命中序列的分数

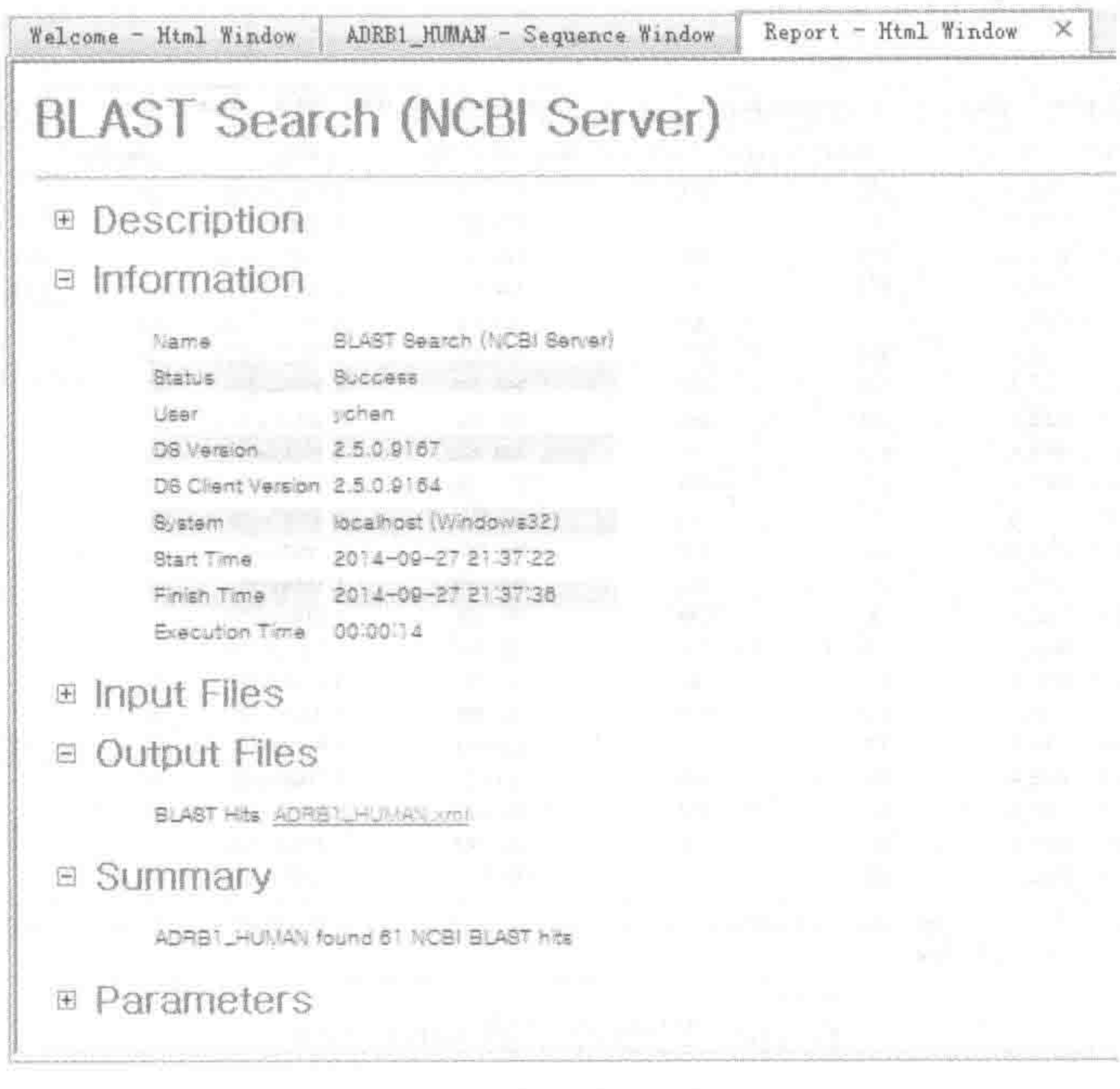

图 1-4 BLAST Search 结果报告

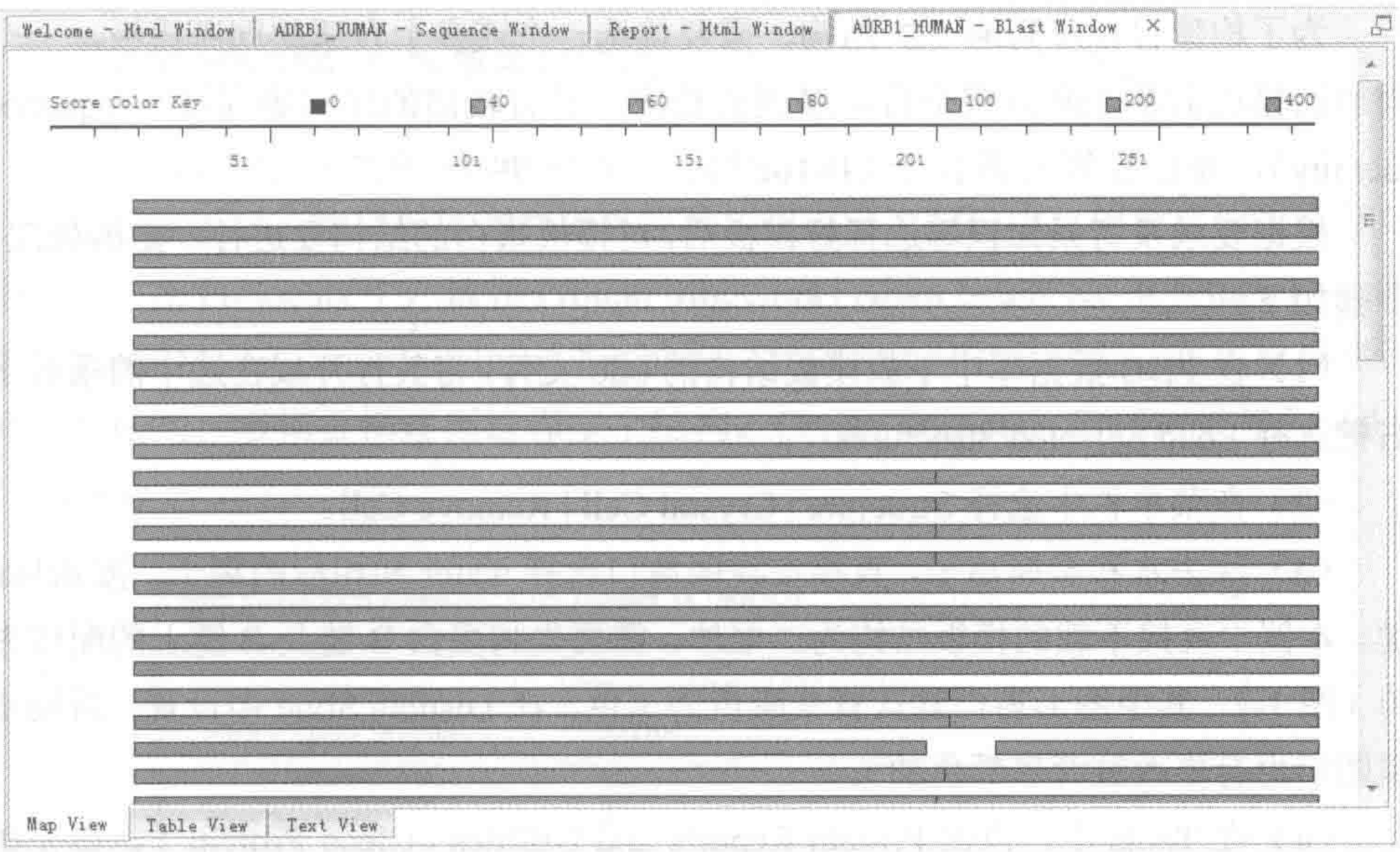

图 1-5 Blast 结果的 Map View（附彩图）

滑动鼠标的中间键可以放大（缩小）Map View 中的结果。

点击窗口下端的 Table View，可显示命中的序列列表（图 1-6），可以看到具体的序列相似性等数值。

Welcome - Html Window | ADRB1_HUMAN - Sequence Window | Report - Html Window | ADRB1_HUMAN - Blast Window ×

	Title/Description	Accession	Sequence Lengt	Alignment Lengt	Bit Score	E-value	Identity	Positive
1	Chain A, Turk···	2Y00_A	315	268	405.601	1.8402e-141	78	88
2	Chain A, Turk···	2VT4_A	313	266	403.29	1.20919e-140	76	86
3	Chain A, Ultr···	4BVN_A	315	268	399.053	6.27901e-139	77	87
4	Chain R, Crys···	3SN6_R	514	287	364.77	1.16353e-122	60	72
5	Chain A, Crys···	3KJ6_A	366	287	358.992	1.77995e-122	60	72
6	Chain A, Crys···	2R4R_A	365	287	358.992	2.18801e-122	60	72
7	Chain A, Crys···	2R4S_A	342	287	354.369	4.8743e-121	60	72
8	Chain A, Stru···	4LDE_A	469	266	358.992	5.16527e-121	64	77
9	Chain A, Stru···	4QKX_A	469	266	358.992	5.45267e-121	64	77
10	Chain A, N-te···	4GBR_A	309	266	347.436	1.07607e-118	64	76
11	Chain A, Stru···	3POG_A	501	180	258.07	1.62307e-81	68	80
12	Chain A, High···	2RH1_A	500	180	258.07	2.08938e-81	68	80
13	Chain A, Irre···	3PDS_A	458	180	252.292	1.1564e-79	68	80
14	Chain A, Chol···	3D4S_A	490	180	251.906	3.51691e-79	67	80
15	Chain A, Stru···	4MQS_A	351	278	125.946	5.12523e-33	28	49
16	Chain A, Crys···	3VG9_A	326	286	123.25	3.61852e-32	34	50
17	Chain A, Ther···	2YDO_A	325	286	119.783	7.07029e-31	33	50
18	Chain A, Ther···	3PWH_A	329	288	118.627	1.61676e-30	33	49
19	Chain A, Crys···	4IAR_A	401	183	113.62	1.76697e-28	35	52
20	Chain A, Crys···	4IAQ_A	403	183	113.62	1.86956e-28	35	52
21	Chain A, Stru···	4DAJ_A	479	179	112.464	7.78387e-28	32	56
22	Chain A, Stru···	3PBL_A	481	186	99.7525	2.52824e-23	33	54

Map View | Table View | Text View

图 1-6　Blast 结果的 Table View

3. 选择合适的模板

为了构建目标序列的 3D 结构，需要挑选一个或多个合适的同源模板。一个理想的模板需要能涵盖整个目标序列的长度，具有较高的序列等同性（sequence identity），并且 E 值要够小（$<1\times10e^{-5}$）。

根据要求及背景知识等选择好模板后，对模板蛋白的结构要进行一定的处理，步骤如下。

（1）在 PDB 数据库中下载模板结构的 pdb 文件，将其打开或在选中的条目上右键选择 Load Selected Structures。

（2）在菜单栏中选择 Structure | Crystal Cell | Remove Cell。

（3）除去水和金属离子，直接在视图窗口选择 water 和相应的离子，按 delete 键。本例中直接下载的模板结构为二聚体，需要先将蛋白 B 链及 B 链上的配体去掉（图 1-7，其中图形窗口默认背景颜色为黑色，在 Display Style 中设置，后续步骤图示也有更改过背景颜色的）。

（4）在 Tools 中，点击 Protein Reports and Utilities | Clean Protein（需事先在 Edit | Preferences 下找到 Protein Utilities | Clean Protein 进行勾选设置）。

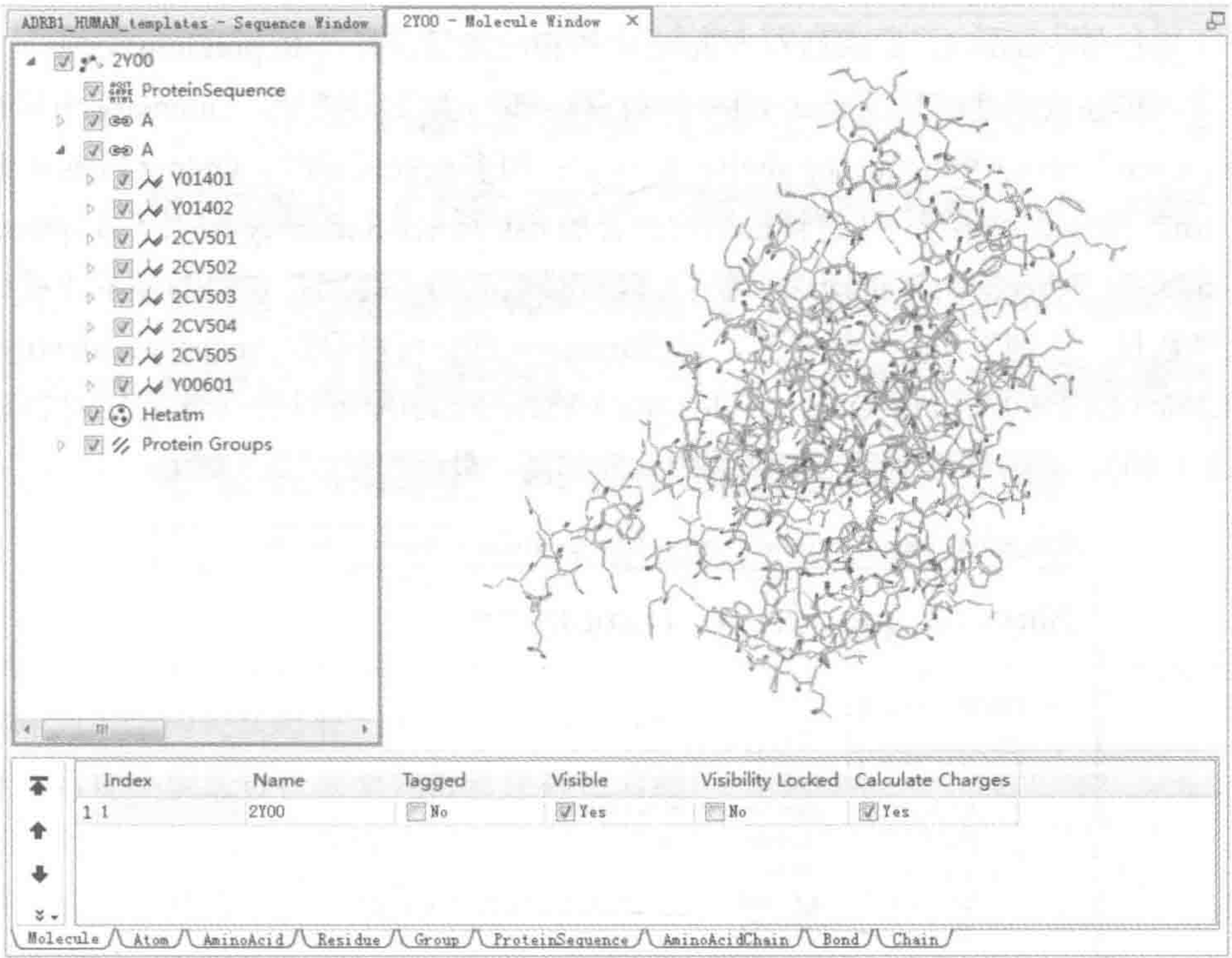

图 1-7　模板结构窗口

二、将目标序列与模板进行比对

1. Align Sequence to Templates

选好模板后，将目标序列和模板序列放在同一个窗口中，模板结构也在窗口中打开。在 Protocols 中，选择 Protein Modeling | Align Sequence to Templates，参数中 Input Model Sequence, Input Templates Structures 分别选择好，单模板模建选择一个结构，多模板模建则选择多个结构。Create Sequence Profile 设置为 False（图 1-8）。

BLAST Search (NCBI Server)　Align Sequence to Templates

Parameter Name	Parameter Value
Input Model Sequence	ADRB1_HUMAN:ADRB1_HUMAN
Input Template Structures	2Y00:2Y00
Create Sequence Profile	False
Align Structures	True

图 1-8　Align Sequence to Templates 参数设置

运行该 protocol，完成后双击 Jobs 中的相应条目，打开 Report.htm 文件。

用单模板模建时，Output Files 中只有 bsml 序列比对结果，Summary 中只有 Sequence identity 和 Sequence similarity 结果；用多模板模建时，Output Files 中除了 bsml 序列比对结果，还有模板结构叠合的 dsv 结果，Summary 中除了 Sequence identity 和 Sequence similarity 结果外，还有模板间的主链间的 RMSD 值和重叠的残基数目。此例采用单模板模建，在 Summary 部分可以看到 Sequence identity= 74.3%（图 1-9），点击 Output Files 中的 View Results 可以打开序列比对窗口（图 1-10），其中颜色越深，氨基酸相似性越高，最深的是完全一致的。

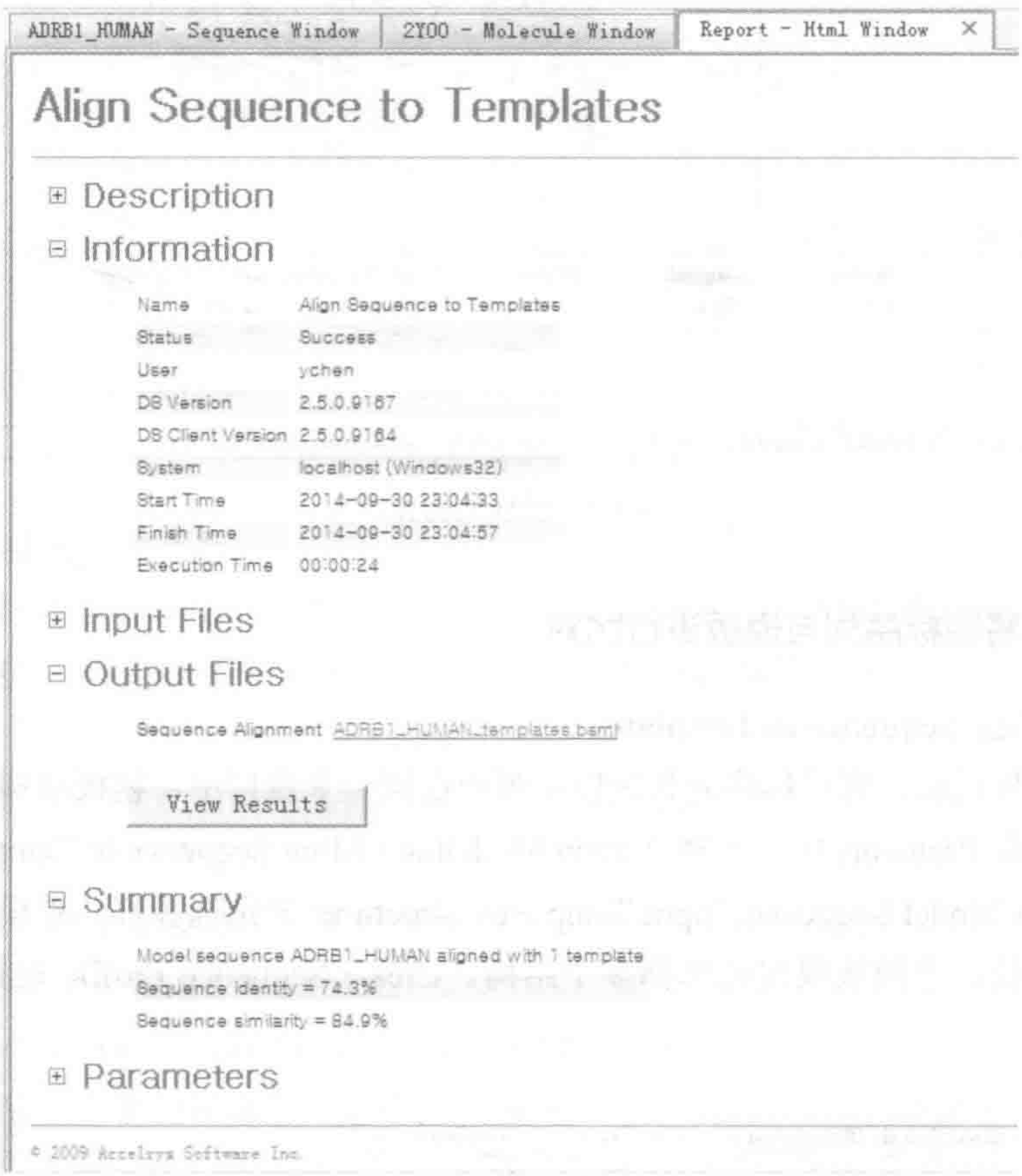

图 1-9　序列比对结果报告

2. Link Sequence and Structure

打开比对后的序列文件和模板结构（多模板模建打开的结构文件为上步中叠合的结构文件），序列窗口为活动窗口时，选中模板序列，在菜单栏中选中 Sequence |

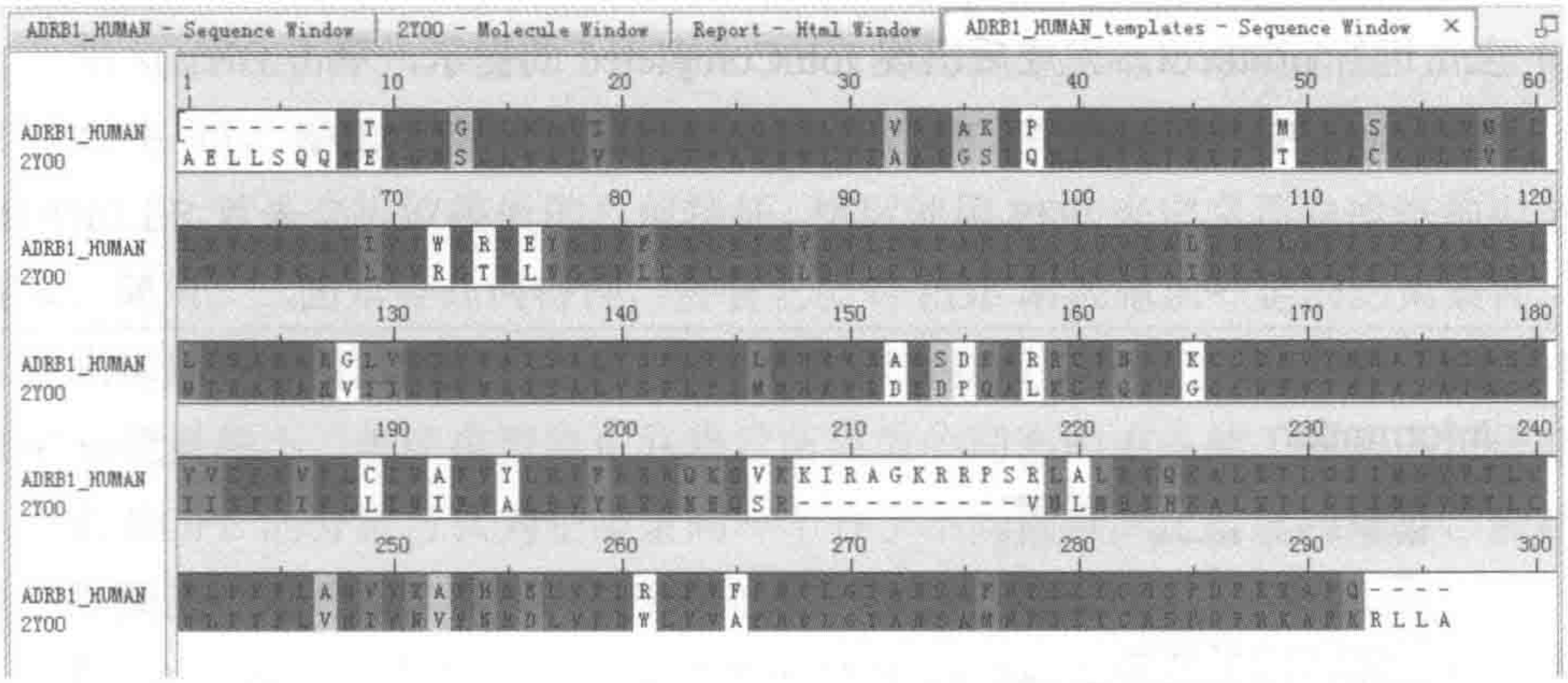

图 1-10　序列比对结果（附彩图）

Link Sequence and Structure…，将模板的序列和相应的结构关联起来（图 1-11）。这个操作开始可能不能进行，需要先打开序列比对结果，将模板序列命名先改为其他，然后打开结构文件，再将模板序列中的命名改为与模板结构中一致的名称。

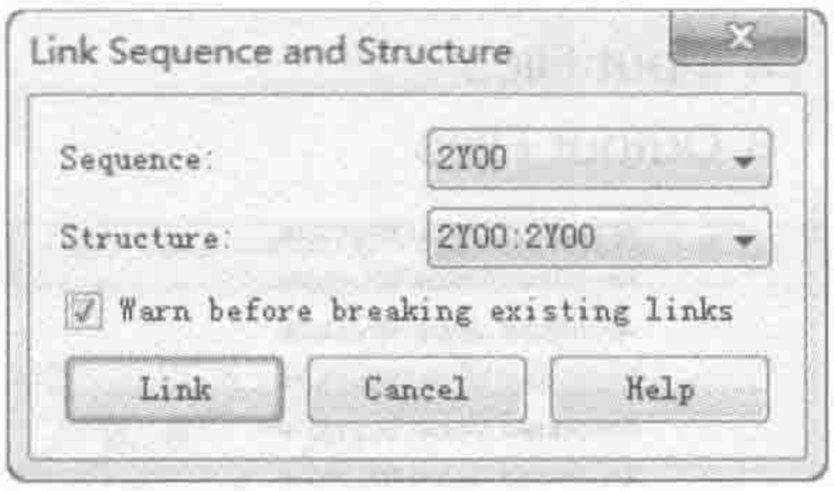

图 1-11　Link Sequence and Structure

三、构建目标序列的三维模型

在 Protocols 中，选择 Protein Modeling |Build Homology Models，在参数设置中，Input Sequence Alignment 选择序列比对后的结果，Copy Ligands 中选择所有的配体，Number of Models 中数目设置为 10，其他参数默认（图 1-12）。当然，可以根据情况设置各种参数，包括二硫键的位置设置，Optimization Level 等。

图 1-12　Build Homology Models 参数设置

点击运行 protocol，完成后出现 Job Completed 的提示，单击 OK。

双击 Jobs 中的相应条目，打开 Report.htm 文件（图 1-13）。

Build Homology Models

⊞ Description

⊟ Information

Name	Build Homology Models
Status	Success
User	ychen
DS Version	2.5.0.9167
DS Client Version	2.5.0.9164
System	localhost (Windows32)
Start Time	2014-09-30 23:25:28
Finish Time	2014-09-30 23:38:54
Execution Time	00:13:28

⊞ Input Files

⊟ Output Files

Annotated Model Structure	ADRB1_HUMAN.B99990001.dsv
Annotated Model Structure	ADRB1_HUMAN.B99990002.dsv
Annotated Model Structure	ADRB1_HUMAN.B99990003.dsv
Annotated Model Structure	ADRB1_HUMAN.B99990004.dsv
Annotated Model Structure	ADRB1_HUMAN.B99990005.dsv
Annotated Model Structure	ADRB1_HUMAN.B99990006.dsv
Annotated Model Structure	ADRB1_HUMAN.B99990007.dsv
Annotated Model Structure	ADRB1_HUMAN.B99990008.dsv
Annotated Model Structure	ADRB1_HUMAN.B99990009.dsv
Annotated Model Structure	ADRB1_HUMAN.B99990010.dsv
Best Model Structure Superimposed to Templates	ADRB1_HUMAN.dsv
Sequence Alignment of Models to Templates	ADRB1_HUMAN_templates.baml

View Results

⊟ Summary

Modeler Version : 9v4

Models Sorted by PDF Total Energy

Model Name	PDF Total Energy	PDF Physical Energy	DOPE Score
ADRB1_HUMAN.B99990005	1341.04	773.72	-38019.61
ADRB1_HUMAN.B99990006	1383.31	807.49	-38645.19
ADRB1_HUMAN.B99990010	1398.97	780.51	-38529.78
ADRB1_HUMAN.B99990009	1400.95	754.32	-38645.59
ADRB1_HUMAN.B99990003	1416.03	747.53	-37998.47
ADRB1_HUMAN.B99990008	1427.52	773.94	-38648.43
ADRB1_HUMAN.B99990007	1444.57	772.02	-38281.45
ADRB1_HUMAN.B99990001	1473.33	852.45	-38249.85
ADRB1_HUMAN.B99990004	1540.33	833.73	-38368.41
ADRB1_HUMAN.B99990002	1615	836.47	-38519.75

⊞ Parameters

图 1-13　Build Homology Models 结果报告

在 Summary 部分，列出了所建立 10 个结构模型的概率密度函数（PDF 值）和 DOPE 值，根据 PDF 总能量值和 DOPE 值进行相应的排序。建模时，DS MODELER 首先会提取模板的几何特征，然后使用 PDF 来定义蛋白质结构中诸如键长、键角、二面角等几何特性，接着它会对 PDF 函数施加一定的约束条件，并以此来构建目标蛋白的 3D 结构，PDF 值可以直接反应所构建模型的好坏。一般，PDF 总能量越小，表明模型能更好满足所提取的同源约束条件，模型的可信度越大。而 DOPE 的分数可认为是衡量同一个分子不同构象的可信度的标准，能够帮助选择预测结构的最优模型，分数越低，可认为模型越可靠。

从 10 个模型中选出一个模型为 PDF 值和 DOPE 值均最小的模型，即能量最小的模型。此例中为第五个模型，在 Output Files 中点击打开此能量最小的模型的文件 ADRB1_HUMAN.B99990005.dsv（图 1-14）。

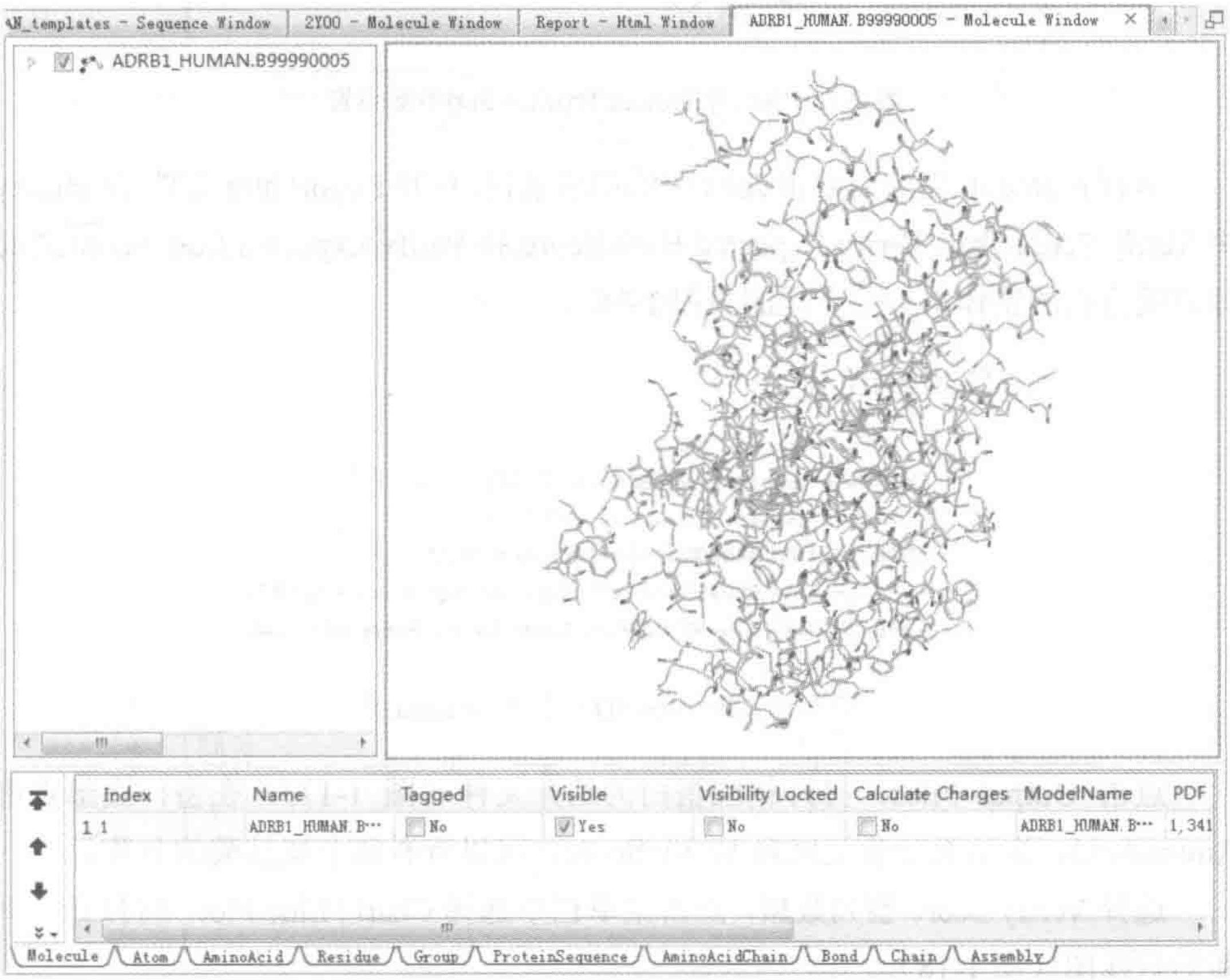

图 1-14　模建得到的能量最小的结构