技术简介

蛋白质和核酸(DNA/RNA)相互依存,共同在各项生命活动中起着重要作用,如 DNA 复制,RNA 转录,RNA 剪接,核酸降解和蛋白质的合成等。蛋白质-核酸相互作用如果发生紊乱,将对细胞的代谢活动以及物种的生存有着重要影响。例如神经系统退行性疾病,还有可能引起癌症。因此,进一步研究蛋白质-核酸相互作用,对寻找解决这些复杂疾病的方法是有利的。本系统首次采用一种基于小样本的学习(Few-shot Learning,FSL)方法来研究蛋白质核酸相互作用界面热点残基预测的问题。首先,从丙氨酸诱变效应数据库中通过处理获得64 个蛋白质核酸复合物。接下来,分别从网状特征,溶剂可达表面积特征,序列特征以及结构特征进行初步的特征提取,一共得到114 维特征。然后为避免特征过于冗余,采用 LassoCV(least absolute shrinkage and selection operator cross-validation)进行二次的特征选择,获得最佳特征子集。然后,利用 TensorFlow深度学习库构建并训练 LSTM(Long Short-Term Memory)模型。最终将训练好的模型部署到网站上,方便用户进行热点残基预测。本系统的网站是利用 Flask和 Bootstrap 框架实现的。用户与平台的前端进行交互操作,并将必要的信息传输给后台,后台对应的模块进行业务逻辑处理。

软件简介

蛋白质和核酸相互依存,相互作用,是支撑一切生命活动的基础。因此,进一步研究蛋白质核酸相互作用,对于了解蛋白质和核酸的生物学功能以及药物研究有着积极的作用。然而,针对目前国内外蛋白质核酸热点预测研究存在的样本量少,基于能量学的预测方法效率低下等问题。本研究首次采用一种基于小样本的学习(Few-shot Learning,FSL)方法来研究蛋白质核酸相互作用界面热点残基预测的问题。并且,为了方便用户使用,构建了可视化操作的网站,实时反馈给用户预测结果,为探索蛋白质-核酸相互作用的本质和为药物的研究提供新的思路和方向。

关键流程设计

下面的流程图描述了整个应用的体系架构。在对用户进行注册后,登录系统,接下来就可以上传需要预测的蛋白质-核酸相互作用残基数据,通过训练好的 LSTM 模型进行预测,即可跳转到预测结果页面。流程如图 1 所示。

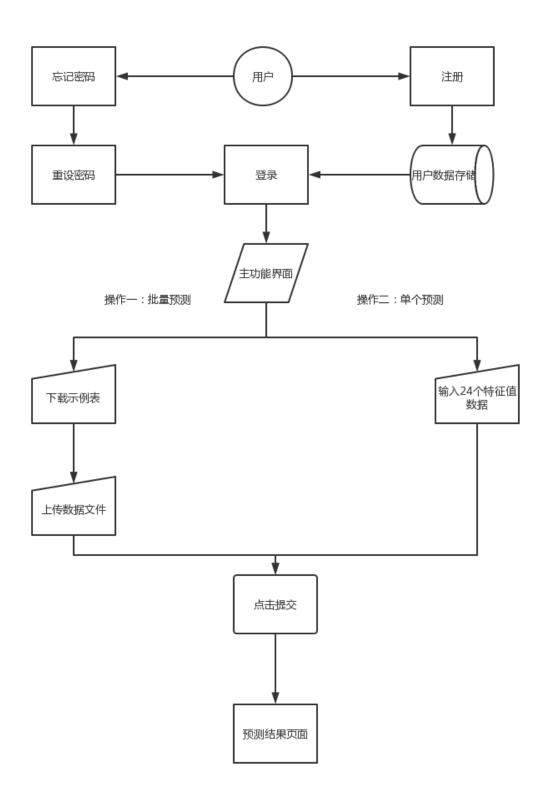


图 1 系统流程图

采用数据访问对象 DAO (Data Access Object)模式。下图阐述了 DAO 对象模型:

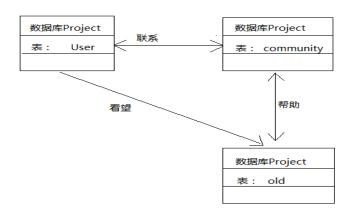


图 2 数据访问对象 DAO

核心算法 LSTM 模型简介

算法构建流程说明:

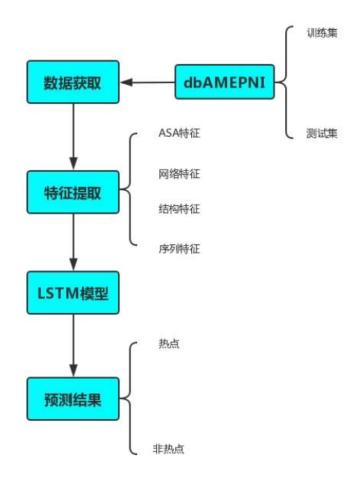


图 3 模型构建流程图

我们需要先获取数据,然后对数据进行清洗、规范化处理,之后利用随机算法分割成训练集与测试集,再将模型搭建好。这样,我们利用训练集训练我们的模型,获取每个节点在最小损失情况下所能达到的当前情况下的最优数据。再利用我们的测试集进行测试,设置例如准确率等指标来衡量我们的模型,并加以优化模型,最终达到理想的效果。

数据集说明:

表 1 数据集分布表

•	数据集		样本总数	正样本	本 负样本 非	热点残	热点残基/非热点残基						
	ij	练	集		150	64	88		0.	705			
	狈	則试	集		64	26	38		0.	684			
Eigenvect 0.528366 0.485684	All-atomsAl 7.9 20.6	1-atomsTo 44.59 37.92	tal-SicT 25.9 21.2	otal-Sicd 38.07 37.45	28. 2 0. 518437 0. 807351 0. 4715	1-Sp-p-All-sd-p-All-sp-p-All-sd-p-All- 94 0.806854 0.781564 0.348669 0.781163 28 0.526055 0.563666 0.475488 0.562633	0 1.081496 1.19351	. 329854 1. 242206 -30	-Total-Sp 30.9 34.6	-All-atcp-All-atcP 10.64 0.426456 34.07 0.518211	SI SF 10.64 28.99	IDER2_FD 6.2 19.1	DDG -0.8
0. 820821	0.5	10, 93	7.8	10, 93		28 0. 525055 0. 563666 0. 475488 0. 562633 84 0. 95122 0. 957093 0. 516048 0. 954955			48. 2	0.51 0.438823	0, 49	0.4	-0.8
0.055122	39	1.64	0.8	1.13		92 0.022727 0.017444 -0.73073 0.017632		0.109139 0.069971 -32.6	28.8	69 -0.74444	63.65	34. 4	-2.
0.477549	5.8	10.49	6.1	10.49		03 0.575472 0.57511 -0.03969 0.57037			46.2	7.75 -0.08377	7.75	4.5	-3
0.027795		159.94	80.2	152.06		32 0.832814 0.841692 0.745931 0.841962			30.1	32.11 0.926404	28.6	16.1	
0.517668	36. 8 73. 4	69.6	34.7 16.1	68, 55 32, 28		05 0.533846 0.532717 0.057584 0.532995 98 0.211286 0.212117 -0.56514 0.212446			14. 2 16. 9	60. 91 0. 250887 120. 7 -0. 59629	60.13 119.9	30.3 60.1	0.
0. 291372	64.9	24.14	10.1	24, 14		63 0.155864 0.155994 -0.31781 0.156047			12.6	130. 61 -0. 24145	130, 61	54.7	0.
0.337118	14.1	28, 26	11.9	24, 69	12, 2 0, 095446 0, 474241 0, 069				13	31, 33 0, 080379	28, 32	13.1	0.
0.429365	71.2	13.75	5. 7	13.75		61 0.081545 0.087535 -0.58835 0.088348			14.4	153. 22 -0. 52347	143.33	64. 2	0.
0.780197	31	99.9	41.9	99.91		26 0.600287 0.615702 0.476457 0.615385			15. 2	66.66 0.715952	62.36	27.9	1.
0.865124	32	58.32	24. 4	58.32		95 0. 471042 0. 475035 0. 265064 0. 47541			16.1	65, 35 0, 285293	64. 45	27.4	0.
0.30547	17.7	18.64	7.8	18.64	9.3 -0.02888 0.31391 -0.036				12.2	40.74 -0.01293	35. 63	17.1	0.
0.401415	34.7	72.86	30.5	72.04		01 0.508333 0.507753 -0.0336 0.507801			14.3	70.34 0.093291	69.84	29.5	0.
0.463416	18.6 5.8	68 28, 5	39. 4 15. 5	65. 32 28. 5		05 0, 713768 0, 722327 0, 071392 0, 722802 25 0, 767327 0, 769854 0, 340746 0, 769841	0 1.629339 2.35628 2 1 1.05353 1.267069 2		24. 3 26. 8	27.13 0.189526 8,52 0.405102	25. 11 8. 52	15.8	0.
0.52517	4.8	56, 06	28	56, 06		25 0.767327 0.769854 0.340746 0.769841 04 0.877743 0.876485 0.272766 0.877551			27. 4	7, 9 0, 527793	7.9	4.7	1.
0. 250325	12.7	87. 52	43.6	87. 52		36 0.804428 0.809022 0.601087 0.808157			24. 2	21.31 0.676837	20.66	10.6	0.
0. 522226	40.7	2, 43	2.1	0	0 -1.14269 0.038058 -1.161			0.115098 0.143101 -27.3	17. 9	61. 42 -1. 08934	31.76	52. 7	1.
0.298351	3.9	57.55	27	48.51		84 0.846395 0.874527 0.771251 0.875399			30	10.42 1.187796	6.96	4.9	1.
1	1.4	15.54	14.4	13.83	19.9 0.227571 0.865738 0.3144	11 0.86747 0.936357 0.377788 0.934272	0 0.686601 0.465108 0	0.511465 0.67978 -46.9	27.1	2.41 0.232978	0.94	2. 2	(
0.03702		136.98	68.6	132.41		86 0.826506 0.822065 0.803699 0.821611			33.1	28.66 1.002491	28.66	14. 4	(
0.128995	50.1	62.03	30.9	53.31		99 0.420981 0.394713 -1.20509 0.394196			18.9	85.31 -1.53251	81.75	42.5	
0.051817	54.5	54. 32	27.1	50.95		24 0.371742 0.364163 -2.23279 0.364061			28.5	92.04 -1.13203	88. 96	45.8	0
0.05354	52. 7 17. 3	57. 53 12. 8	28.6	57. 53 12. 8		59 0, 351784 0, 400571 0, 236799 0, 401136 37 0, 308824 0, 311891 0, 02086 0, 310757			24 22.1	105, 77 0, 181607 28, 24 0, 023104	86. 09 28. 24	52. 7 14. 1	(
0.835295	13. 4	63	31. 4	63		23 0, 742317 0, 742487 0, 711586 0, 742308			23, 8	21, 85 0, 848928	21, 85	10.9	0
0.457154	20.4	40.69	20. 2	40.38		42 0.544474 0.548046 0.434862 0.547672			17.9	33. 91 0. 406742	33. 3	16.9	0
0.504853	18	65. 66	36.7	57.5		16 0.682156 0.693774 -2.1628 0.693356			21.8	30.51 -2.1826	25. 38	17.1	0
0.506602	0	75. 98	39.1	75.78	48. 4 1. 284531 1 1. 0522	88 1 1 1.012552 1	0 1.258076 1.661842	. 603679 1. 320309 -45. 2	28.6	0 1.341782	0	0	0
0.371818	45. 6	23.42	16.3	23.42		53 0.321499 0.326002 -1.0548 0.325444			22.5	49.55 -0.78515	48. 42	34. 4	0
0.191757		148.82	62.3	148.35		14 0.877465 0.916306 1.542602 0.916667			19.8	20.74 1.729103	13, 55	8.7	0
0.262704	31.3	77.02	32. 3	76.84		59 0.550256 0.549839 0.166933 0.548991			26	63.1 0.242865	62. 91	26.4	1.
0.154663	47. 4	60.75 77.52	25. 4	60.75 76.41	30. 2 0. 02287 0. 387165 0. 0977 43. 1 -2, 24795 0. 977184 -1, 878	12 0.386606 0.388924 0.036492 0.389175			24.4	96.16 -0.03128	95. 45	40.3	0
0.412323	3, 5	27	12.6	76. 41		34 0. 978552 0. 97686 -1. 88261 0. 977324 33 0. 807692 0. 811298 0. 155658 0. 81383			22. 7	1.81 -2.24214 6.28 0.112559	1.81 6.28	0.8	0
0.700419	12.6	19, 56	9. 2	19.56		34 0, 467005 0, 466158 0, 172539 0, 468354			13.8	22. 4 0. 178169	22. 4	10.5	0
	46.0	20	11.6	19.99		99 0.666667 0.691696 0.317281 0.693023			37.3	10.03 0.250296	8. 91	5.8	2

图 4 部分数据集展示

模型性能评价:

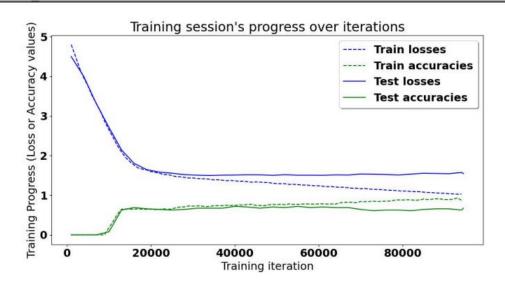


图 5 训练集的迭代过程

本系统的 LSTM 模型的各项评价指标为: ACC= 72% , AUC= 70% , F1_score=60% 。

常用模型性能对比:

模型	ACC	AUC
RF	0.65	0.62
KNN	0.62	0.57
SVM	0.61	0.60
LSTM	0.72	0.70

表 2 不同分类模型在测试集上的性能比较

表 2 展示的是四种不同的分类模型在独立测试集上的 ACC, AUC 比较。结果表明, LSTM 模型的 AUC=0.70, ACC=0.72 明显高于其他三类模型。因此, 本研究选择 LSTM 模型构建预测模型, 从而获得最优的结果。

软件主要功能简介

该软件主要包含以下功能:

- 注册登录模块;
- 批量上传样本数据模块;
- 单个样本数据输入模块;

● 预测结果返回模块;

注册登录模块

登录原型页面说明:

页面结构布局:

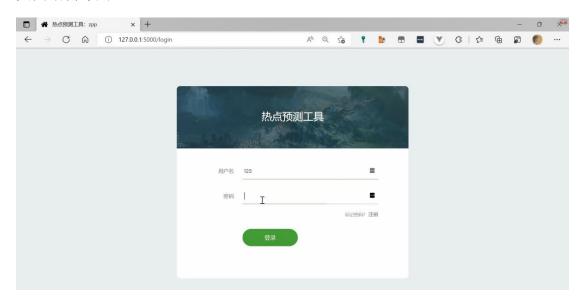


图 6 主操作页面

页面结构:显示用户名,密码输入框,输入框右下角有"忘记密码?"和注册跳转链接。

注册原型页面说明:

◆ 页面结构布局:

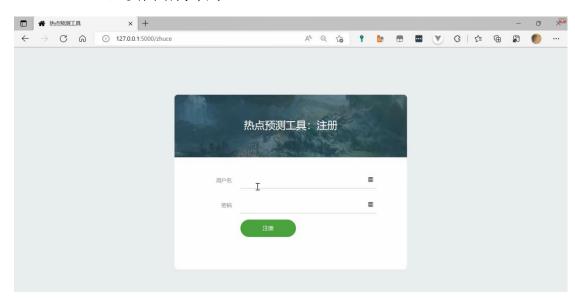


图 7 注册页面

- 页面结构: 自上而下依次设置: 用户名,密码注册信息。 最后显示"注册"。
- 1. 输入用户名、密码,点击注册按钮。需要验证以下内容:
- 验证是否都不为空,若有空的则显示对话框"xx 不能空!"。
- 验证用户名是否已经存在,若已经存在,则弹出对话框 "用户名已存在,请重新输入!"
- 以上全部验证通过,则将数据存入数据库,并直接跳入登录页面。

主页面模块

- 1. 该页面为主页面,点击对应的链接,即可进入对应的页面。
- 2. 点击"样本数据.csv",跳转至下载示例文件页面。
- 3. 点击"选择文件"按钮, 跳转批量上传文件页面。
- 4. 输入24个特征值数据,点击提交跳转预测结果页面。



12	p-p-All-atoms-REL	Interfacial neighborhood property 1 based on total relative ASA	0
13	d-p-All-atoms-REL	Interfacial neighborhood property 2 based on change in total relative ASA upon complexation	0.64
4	p-p-Total-Side-ABS	Interfacial neighborhood property 1 based on side-chain absolute ASA	0
15	d-p-Total-Side-ABS	Interfacial neighborhood property 2 based on change in side-chain absolute ASA upon complexation	1
6	p-p-Total-Side-REL	Interfacial neighborhood property 1 based on side-chain relative ASA	1
7	d-p-Total-Side-REL	Interfacial neighborhood property 2 based on change in side-chain relative ASA upon complexation	1
8	donor-num	The number of hydrogen bond donor residues	1
9	p-Total-Side-REL	Relative change in side-chain relative ASA upon complexation	-8.6
20	p-Total-Side-ABS	Relative change in side-chain absolute ASA upon complexation	38.8
11	p-All-atoms-ABS	Relative change in total absolute ASA upon complexation	21.7
22	p-All-atoms-REL	Relative change in total relative ASA upon complexation	0
23	PSI	IUPAC peptide backbone torsion angles PSI	15.13
14	SPIDER2_hsb_CN	Half-sphere Cα-Cβ contact numbers	9-7

图 8 数据提交页面

批量上传样本数据模块

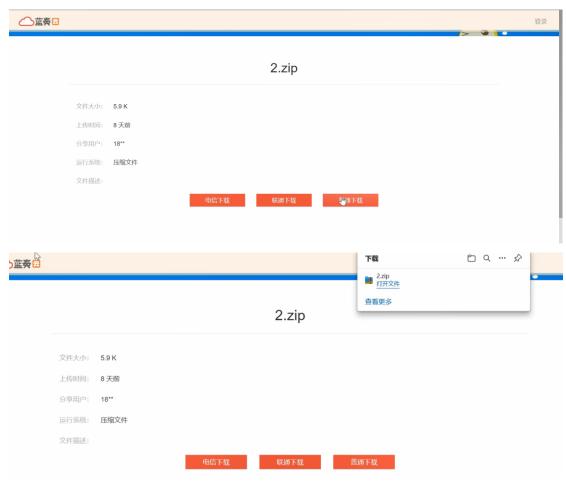


图 9 示例文件下载页面

示例文件存储在蓝凑云中,点击普通下载即可下载示例文件。

返回主操作页面,点击"选择文件"上传和示例文件格式一致的 CSV 文件,点击"提交"按钮即可跳转至批量样本预测结果页面。"1"表示该样本是热点残基,"0"表示该样本是非热点残基。

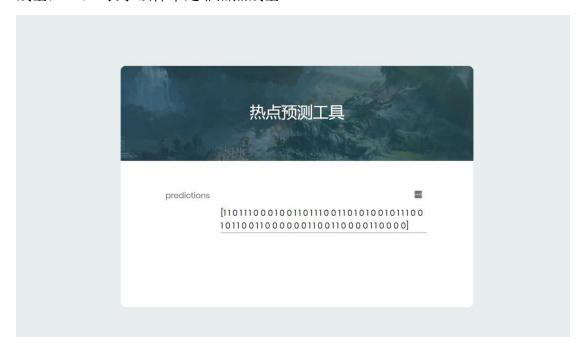
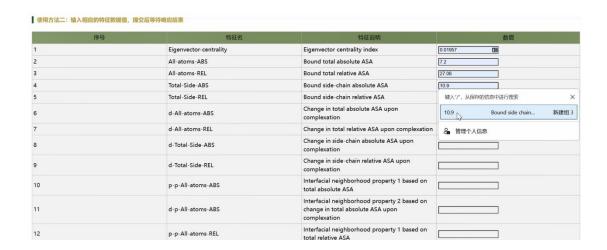


图 10 批量预测结果页面

单个样本数据输入模块



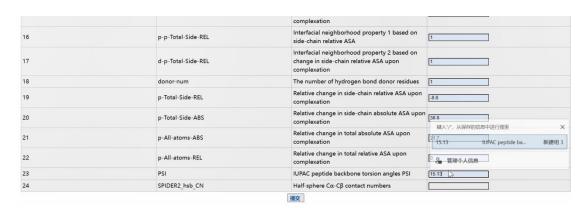


图 11 单个样本特征值提交页面

用户需要根据特征名,输入对应的特征值数据。输入完毕后点击提交按钮,即可跳转至单个样本预测结果页面,如下图所示:

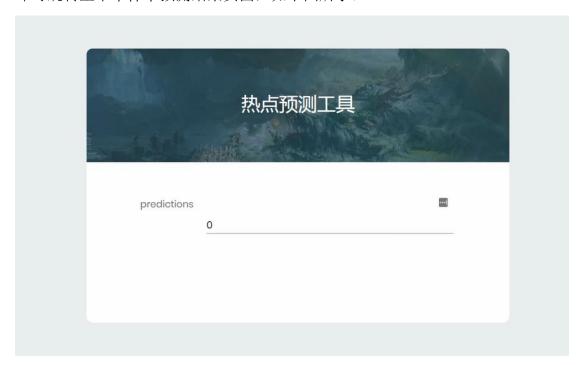


图 12 单个样本预测结果页面