
Homology Modeling of CEA

—Protein Tertiary Structure Prediction

Wang Jilong

2008-5-19

结构预测的作用

- 预测的结构信息虽没有X-射线或NMR方法得到的蛋白质结构那么准确，但预测得到的结构信息可以用来合理设计生化实验，如：点突变，蛋白质的稳定性或功能分析。另外，预测的结构信息也可以用来辅助分析一些实验结果。
-

Homology Modeling

■ 基本原理

两个蛋白质的序列具有很高的相似性，则它们可能具有相似的三维结构。

如果通过X-射线或NMR方法得到一个蛋白质（**template**）的结构，就可以将该结构复制给另一个与其序列相似的结构未知蛋白，由此预测的蛋白质三维结构具有很高的可信度。

同源建模通过未知结构蛋白与**template**比对，可以产生一个全原子坐标结构模型。

Homology Modeling

基本步骤：

- 挑选模板（Template Selection）；
 - 序列比对（Sequence Alignment）；
 - 基本骨架模型的建立 (Backbone Model Building)；
 - 连接环的建模 (Loop Modeling)；
 - 侧链精修 (Side Chain Refinement)；
 - 使用能量方程进行模型精修 (Modeling Refinement Using Energy Function)。
 - 模型评估（Model Evaluation）
-

Homology Modeling

一·挑选模板 (template selection)

- 在蛋白质结构数据库中挑选合适的同源序列来作为建模的模板，通常在PDB数据库中搜索。

这种搜索可以使用BLAST或FASTA等启发式序列比对搜索程序。也可以使用SSEARCH或ScanPS等动态规划法的搜索程序，这样可以得到更灵敏的搜索结果，而相对较小的蛋白质结构数据库，是这种动态规划法搜索得以实施的基础。

Homology Modeling

一.挑选模板 (template selection)

- CEA是免疫球蛋白家族成员，具有七个免疫球蛋白结构域。这里主要对其氮端三个结构域CEA(N), CEA(1A), CEA(1B) 进行建模。挑选了免疫球蛋白家族成员CD2(R1), CD4(H1), CD4(H2), 1REI(V), 3FAB(VH), 和3FAB (VL) 作为待选模板，然后使用ALIGN程序进行序列比对，来选择合适的模板蛋白序列。

Sequence comparisons of CEA domains N, IA and IB with other Ig-folds

	CD2 (R1)	CD4 (H1)	CD4 (H2)	1REI (V)	3FAB (VH)	3FAB (VL)
CEA(N)	3.11	2.98	0.41	4.09	3.32	0.77
CEA(IA)	2.45	2.24	2.28	4.53	2.81	2.97
CEA(1B)	3.85	2.40	4.19	2.49	2.74	3.75

这里选择CD4(H2)作为CEA (1B) 的主要模板，是因为它们具有较高的相似性，但它没有该保守二硫键，故还选择CD4(H1)和1REI(V)作为模板构建二硫键 (C225-C265) 对应区域，即225-237位和252-265位使用CD4 (H1) 为模板。最终CEA(N)以CD2(R1), CD4(H1) 和1REI(V) 为模板；CEA(1A) 和CEA(1B) 以CD4(H1), CD4(H2) 和1REI(V)为模板。

Homology Modeling

二.序列重新比对-建模中最关键的一步

目标蛋白和模板蛋白需要使用修改过的比对算法进行重新比对，来找到最佳比对结果。

因此，应该**综合使用不同比对算法的程序**，如**Praline**和**T-Coffee**等，进行序列比对。但是最好的比对算法也可能出错，所以我们应该**人工检查**，来保证保守的关键残基进行了正确的比对。必要时，还要做一些手动修改来提高比对质量。

Homology Modeling

二.序列重新比对-建模中最关键的一步

在**CEA**结构预测中，使用了手动修改，来确保以下几方面：

- 保证关键的折叠标记位点进行了比对，如**C**链保守的色氨酸，保守的盐桥和二硫键连接；
 - 保证每个折叠中保守的疏水核心比对上；
 - 保持 β 链的结构，即将所有的插入和删除突变都在环形连接中进行；
 - 在可能的地方，以上几方面应该在待建模序列与其它临近的家族蛋白成员的比对中保持一致。
-

Homology Modeling

二.序列重新比对-建模中最关键的一步

		Table II					
		β -strands and loop fragments used in CEA model building					
		Domain	Res. range	Template	Fragment		
1-109	CEA (N)		1-4	REI			
	CD2 (1)		5-8		2HLA(A)		
	CD4 (1)		9-20	CD2(1)			
	REI		21-23		3MCG(1)		
	BURIED		24-35	CD2(1)			
	STRAND		36-42		4FAB(H)		
			43-48	CD4(1)			
110-202	CEA (IA)		49-52		2MCP(L)		
	CD4 (2)		53-58	CD4(1)			
	CD4 (1)		59-63	REI			
	REI		64-68	CD4(1)			
	BURIED		69-70	CD2(1)			
	STRAND		71-75	CD4(1)			
			76-92	CD2(1)			
			93-98		2HFL(H)		
			99-106	CD2(1)			
			107-112		3FAB(L)		
		LINK 1	113-132	CD4(2)			
		CEA(IA)	133-144	CD4(1)			
			145-148	REI			
			149-155		2HFL(L)		
			156-166	CD4(2)			
			167-181	CD4(1)			
			182-185	CD4(2)			
			186-190		2HFL(L)		
			191-199	CD4(2)			
		LINK 2	200-205		3FAB(L)		
		CEA(IB)	206-209	CD4(2)			
			210-211		3HFM(H)		
			212-224	CD4(2)			
			225-237	CD4(1)			
			238-243		1MCP(L)		
			244-245	REI			
			246-251	CD4(2)			
			252-265	CD4(1)			
			266-270	CD4(2)			
			271-275		1MCW(M)		
			276-284	CD4(2)			
		LINK 3	285-290		3FAB(L)		

<pre> 1-109 CEA (N) KLTIE-STPFNVAREGKVEVLLLVH--NLPQHLFGYSHYKGERVDDG CD2 (1) -----SQTWNGALGHGINLNINPQMTDDIDEVRHERG----- CD4 (1) -----KKVVLGKKGDVVELTCT--ASQKKSIQFHWKNS-----NC REI QMTQSP-SSLSASVGDRTITCQ--ASQDIKYLNNHYQQTGPKAE BURIED * * * * * STRAND -----A----- ---B-- --C--- </pre>	<pre> 110-202 CEA (IA) LRPSSISSNNSKPFVEOKDAVAFTCEPETQDATYLNHWVNNQSLPVS CD4 (2) ----GLTANS DTHLLQGQSLTITLESPPGSSPSVQCRSP----- CD4 (1) ----KKVVLGKKGDVVELTCTASQKKS IQFHWKNS----- REI DIQMTQSP-SSLSASVGDRTITCQASQDIKYLNNHYQQTGPG-- BURIED * * * * * STRAND -----A----- ---B-- --C--- </pre>	<pre> 203-287 CEA (IB) PDAPTISP-LNTSYRSGENLNLSCLAAASNPPAQYSHFVN----G CD4 (2) ----GLTANS DTHLLQGQSLTITLESPPGSSPSVQCRSP-----RG CD4 (1) ----KKVVLGKKGDVVELTCTASQKKS IQFHWKNS-----NC REI DIQMTQSP-SSLSASVGDRTITCQASQDIKYLNNHYQQTGPKAP BURIED * * * * * STRAND -----A----- ---B-- --C--- </pre>	<pre> .IQNIIQNDDTGFYTLHVTKSDLVNEEATG-QFRVYPE .IKNLTRDDSGTYNVTVYSTH-GTRILNK-ALDLRIL .IKNLKI EDSDTYICEVE----D---QKE-EVQLLVF .ISSLQPEDIATYYCQYQSLPY---TFGGQTKLQIT * * * * * - - - - -F--- - - - -G--- </pre>
<pre> NITVNNSGSYCCAHNSDTGLNRTTPT-TITVYAE QLQLQDSGTWCTVLRNQ----KKVQF-KIDIVVL NLKI EDSDTYICEVE----DQKE-EVQLLVF SLQPEDIATYYCQYQSLPY---TFGGQTKLQIT * * * * * - - - - -F--- - - - -G--- </pre>			

Homology Modeling

三. 基本骨架模型的建立

待建模蛋白与**template**比对上的残基可以认为其与模板的相应区域具有相似结构。这样就可以将模板蛋白对应残基的坐标简单复制给待建模蛋白：

- 如果比对区域的比对**残基相同**，可以将**template**残基的侧链原子同主链原子的坐标同时复制给待建模蛋白。
 - 如果比对区域的比对**残基不同**，仅将**template**骨架原子的坐标复制给待建模蛋白，而待建模蛋白的侧链原子坐标将在后面侧链精修的步骤中进行。
-

Homology Modeling

三. 基本骨架模型的建立

在骨架原一个模板蛋白。但分子。在使用模和叠加。可以选每个模板与待建坐标。CEA构建使用CD2 (R1) 使用了1REI(V)。

Domain	Res. range	Template	Fragment
CEA(N)	1-4	REI	
	5-8		2HLA(A)
	9-20	CD2(1)	
	21-23		3MCG(1)
	24-35	CD2(1)	
	36-42		4FAB(H)
	43-48	CD4(1)	
	49-52		2MCP(L)
	53-58	CD4(1)	
	59-63	REI	
	64-68	CD4(1)	
	69-70	CD2(1)	
	71-75	CD4(1)	
	76-92	CD2(1)	
	93-98		2HFL(H)
	99-106	CD2(1)	
	107-112		3FAB(L)
	LINK 1 CEA(IA)	113-132	CD4(2)
133-144		CD4(1)	
145-148		REI	
149-155			2HFL(L)
156-166		CD4(2)	
167-181		CD4(1)	
182-185		CD4(2)	
186-190			2HFL(L)
191-199		CD4(2)	
200-203			3FAB(L)
206-209		CD4(2)	
210-211			3HFM(H)
LINK 2 CEA(IB)	212-224	CD4(2)	
	225-237	CD4(1)	
	238-243		1FBI(L)
	244-245	REI	
	246-251	CD4(2)	
	252-265	CD4(1)	
	266-270	CD4(2)	
	271-275		1MCW(M)
	276-284	CD4(2)	
	285-290		3FAB(L)
LINK 3			

的方法是只使用一用了三个模板蛋白三个模板进行对比的平均值或者选择构建时使用的原子(N)中：9-20位(H1)，而59-63位

Homology Modeling

四.连接环(loop)的建模 一误差主要来源

在为建模而进行的序列比对中，往往会出现由于插入或缺失造成的空位，这时需要使用连接环建模来填补结构上的空位。这一步也是同源结构建模中非常困难的一步，是同源建模误差的主要来源。

连接环(loop)的建模可以看作是微型蛋白建模问题。至今没有成熟的方法来构建可靠的连接环结构。目前有两种方法来构建连接环(loop)：**数据库搜索法**(database searching method)和**从头构建法**(ab initio method)。

Homology Modeling

四.连接环(loop)的建模 一误差主要来源

➤ 数据库搜索法(database searching method)

从已知蛋白质结构的数据库中找到与待建模蛋白loop环前后的主链原子形成的茎(stem)区域空间结构相匹配的区域。首先测定茎锚定区域的方向和距离，从PDB数据库中搜索具有相同长度，并与上面的端部构象吻合的片段。最佳loop环的选择要考虑序列相似性，并且与临近部分具有最少的立体结构冲突。最佳匹配片段的构象复制给茎的连接区域。

➤ 从头构建法(ab initio method)

可以产生许多随机的loop结构，然后从中挑选一个loop，与邻近侧链不发生冲突，具有较合理的低能量状态，并且 Φ 和 Ψ 角位于允许区域。

➤ 一些专业loop构建软件有：FREAD（使用数据库搜索法）； PETRA（使用从头构建法）；CODA(综合运用两种方法)。

Homology Modeling

四.连接环(loop)的建模 — 误差主要来源

CEA中的loop (Brookhaven数据 2HFL...), 从Link1使用3FAB (H)。

Domain	Res. range	Template	Fragment
CEA(N)	1-4	REI	
	5-8		2HLA(A)
	9-20	CD2(1)	
	21-23		3MCG(1)
	24-35	CD2(1)	
	36-42		4FAB(H)
	43-48	CD4(1)	
	49-52		2MCP(L)
	53-58	CD4(1)	
	59-63	REI	
	64-68	CD4(1)	
	69-70	CD2(1)	
	71-75	CD4(1)	
	76-92	CD2(1)	
	93-98		2HFL(H)
	99-106	CD2(1)	
	107-112		3FAB(L)
	LINK 1 CEA(IA)	113-132	CD4(2)
133-144		CD4(1)	
145-148		REI	
149-155			2HFL(L)
156-166		CD4(2)	
167-181		CD4(1)	
182-185		CD4(2)	
186-190			2HFL(L)
191-199		CD4(2)	
200-203			3FAB(L)
LINK 2 CEA(IB)	206-209	CD4(2)	
	210-211		3HFM(H)
	212-224	CD4(2)	
	225-237	CD4(1)	
	238-243		1FBI(L)
	244-245	REI	
	246-251	CD4(2)	
	252-265	CD4(1)	
	266-270	CD4(2)	
	271-275		1MCW(M)
LINK 3	276-284	CD4(2)	
	285-290		3FAB(L)

索法构建的。从 FAB, 4FAB, 1, 如107-112位的link使用2HFL

Homology Modeling

五.侧链精修

- 完成主链原子（及比对中identity的侧链原子）坐标构建后，需要确定序列比对中那些不同残基位点的侧链原子坐标。
 - 侧链的几何构象在评估蛋白活性位点上的蛋白-底物相互作用以及在蛋白质相互接触面上的蛋白-蛋白相互作用中，有非常重要的作用。
-

Homology Modeling

五·侧链精修

- 理论上侧链原子坐标的确定可以通过搜索侧链在每种扭转角度下能呈现的所有构象，来找到一个与临近原子具有最低相互作用能量构象作为侧链的原子坐标。但这种方法需要繁重的计算机计算工作，受到计算技术发展的限制往往无法实现。
- 当前常用的侧链预测程序使用**旋转异构体**（rotamers）的概念。旋转异构体是偏向于从已知蛋白晶体结构中提取的侧链扭转角度。这些倾向的侧链构象的集合构成了一个旋转异构体文库，并且文库中的旋转异构体按出现的频率排序。这样可以大大减少需要尝试的侧链构象数。预测的侧链构象选择与临近原子具有最低相互作用能量的旋转异构体。
- 很多时候，使用旋转异构体的方法仍然非常费时。为了进一步缩短搜索时间，可以将骨架构象也考虑进去。因为**骨架构象与一些特定的旋转异构体是相对应的**。使用最适旋转异构体构建侧链构象后，还要对其进行优化来减少其与其它区域结构在立体构象上的重叠。
- 预测侧链结构的软件有：**SCWRL**

Homology Modeling

五·侧链精修

CEA侧链结构预测：

- 如果比对中的残基与**CEA**相同，则使用模板等价残基的构象作为**CEA**的残基构象；
 - 对于发生保守取代的，**CEA**的侧链排布成与模板对应残基类似的区域；
 - 对于发生非保守取代的，**CEA**侧链排布成该残基通常观察到的构象；
 - 最后，人工检查将发生严重立体结构冲突的构象用另外可接受的构象替换掉。
-

Homology Modeling

六·使用能量方程进行模型精修

- 在上面连接环（loop）构建，侧链构建的步骤中都使用了势能计算来改善结构模型。但这样并不能保证整个模型不会出现结构异常，如：不利的键角，键长或过密的原子接触。这些结构的异常可以通过在整个模型范围应用能量最小化步骤来矫正，使得整个蛋白构象处于能量势处于最低状态。能量最小化的目的是在不显著改变整体结构的前提下，缓解立体结构的不利碰撞和张力。
- 但是能量最小化在使用时需要特别注意，因为过度的能量最小化可能导致残基离开其正确的位置。因此，推荐有限的能量最小化，来移除主要的错误，如短的键长和紧密的原子冲突。如果需要，关键的保守残基和那些设计辅因子结合的区域需要被限制使用能量最小化。

Homology Modeling

六·使用能量方程进行模型精修

- 另外还有一种结构模型精修的方法是分子动态模拟（molecular dynamic simulation），因为能量最小化原理只能使原子趋于当前结构的最小化，而没有搜索所有可能的构象，结果导致一种次优化结构（suboptimal structure）。寻找全局能量最小化，需要在势能图谱中移动原子，使其能量上升或下降。分子动态模拟法通过加热或冷却来模拟分子能量上升和下降的运动。这样可以克服一些能量最小化原理无法接近的能量垒。人们希望这种模拟可以追踪蛋白质的折叠过程，有一个更好的机会得到蛋白质的真正结构。
 - 目前的结构模型精修软件有：**GROMOS**使用分子模拟法。
 - **CEA**使用**CHARMM**程序来进行结构精修，首先使用**25**步来修改主要的立体结构冲突，然后通过**500**步得到能量最小化。这种最小化得到较好的肽段几何构象并修改了所有的立体结构冲突。最终每个结构域都得到一个稳定的能量值。
-

Homology Modeling

七.模型评估

- 最终的同源结构模型需要通过评估来确保其结构的特性符合其生理生化规律。这包括检验 Φ — Ψ 角度的异常性，键长，紧密接触面等。另一种检验蛋白质结构模型质量的方法是固定性考虑这些立体化学特性。这种方法可通过编辑实验得到的蛋白质结构的空间特性和相互作用能的统计谱，来发现模型的错误。通过比对统计参数与构建的模型，可以知道哪些区域折叠正常，哪些是异常的。
- 常用的评估程序有：
Procheck检测常用生理化学参数；WHAT IF检验化学正确性；ANOLEA使用统计计算的方法；Verify3D使用统计计算的方法。
注意ANOLEA和Verify3D虽然都使用统计方法，但它们的阈值不同，并且分数高低代表的含义也不同：ANOLEA分数越低模型越正确，而Verify3D中分数越高越正确。
- CEA在构建完个结构域的结构后使用POL_DIAGNOSTICS进行模型评估。

Homology Modeling

现在已经出现一些软件，可以自动进行同源建模结构预测：

➤ **Modeller**

需要提供比对的模板序列

➤ **Swiss-Model**和**3D-JIGSAW**

可以全自动建模，也可以选择手动输入比对的模板序列。

另外，现在也出现了一些通过同源建模法预测的蛋白质三维结构数据库：

ModBase和**3Dcrunch** 它们可以提供蛋白质结构进化的一些有用信息。

Homology Modeling

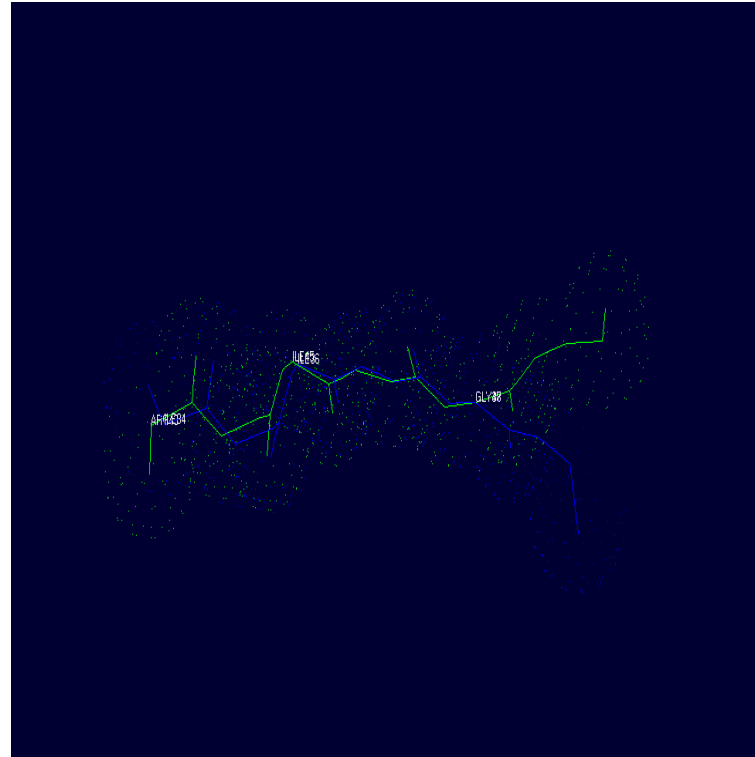
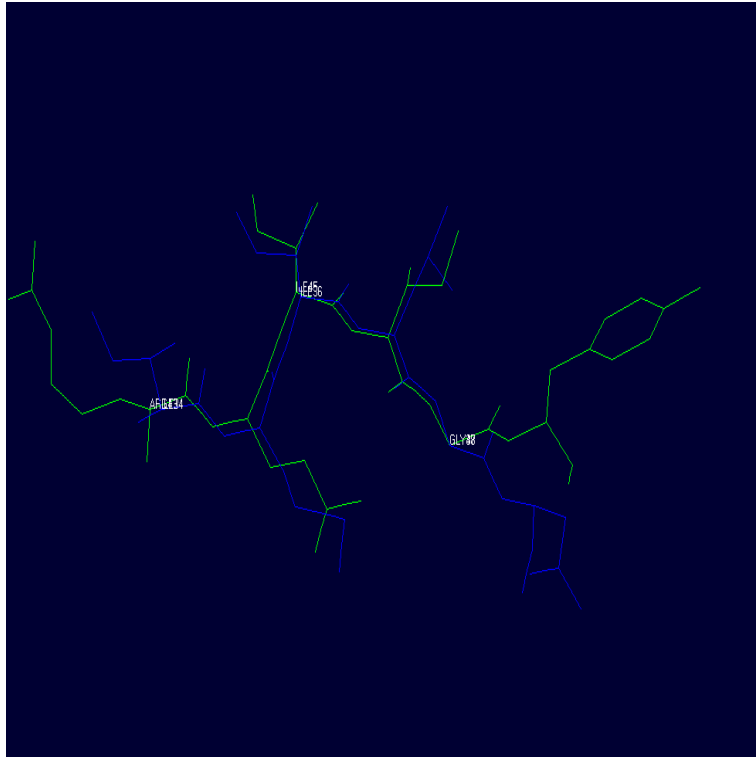
CEA预测结构的一些分析

- CEAM5_HUMAN 的氮端34-110位的X射线晶体结构**2QSQ**
- CD4_HUMAN 第一，二个结构域的X射线晶体结构**3CD4**

由于在CEAM结构预测中43-48残基是使用CD4（H1）作为template，故在Swiss-PdbView中选择2QSQ的43-48位残基**RQIIGY（绿色）**，与其对应的3CD4中应为34-39位残基**IKILGN（蓝色）**，将这两段肽链进行叠合（superimposed），结果如图

Homology Modeling

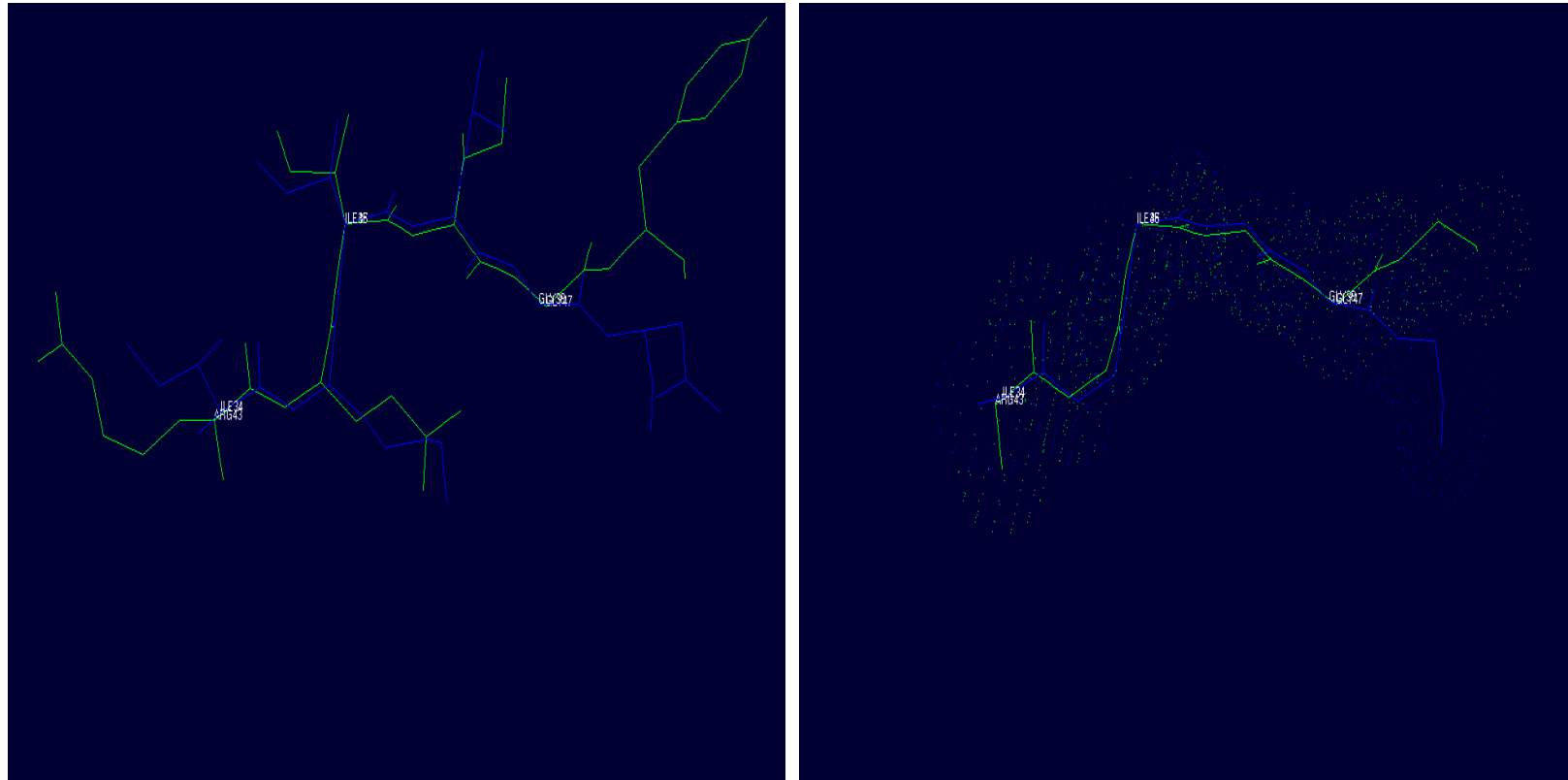
CEA预测结构的一些分析



图一，以CEA-47G与CD4-38G为第一对叠合原子，CEA-45I与CD4-36I为第二对叠合原子，CEA-43R与CD4-34I为第三对叠合原子，右图为左图去除侧链集团。

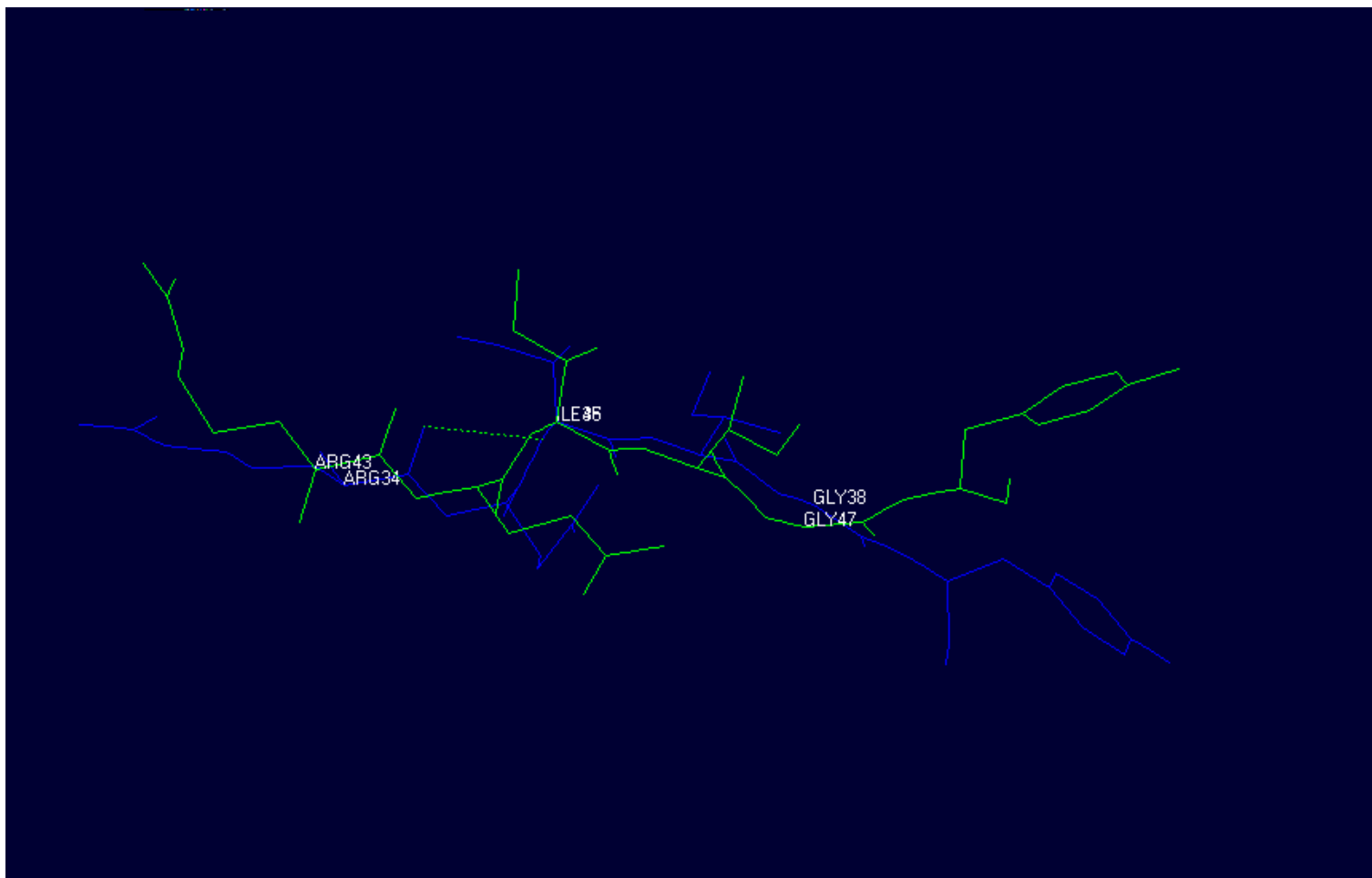
Homology Modeling

CEA预测结构的一些分析



图二，以CEA-45I与CD4-36I为第一对叠合原子，CEA-47G与CD4-38G为第二对叠合原子，CEA-43R与CD4-34I为第三对叠合原子，右图为左图去除侧链集团。

将3CD4中应为34-39位残基IKILGN（蓝色）突变成RQIIGY，再与2QSQ的43-48位残基RQIIGY（绿色）叠合：



蛋白质三维结构预测的其它方法

➤ Threading

将未知蛋白的序列放入结构数据库中，并选择最适合的折叠，以此来预测蛋白质的三维结构。该方法不需要一级序列的相似性。该方法在一个折叠文库中计算一个氨基酸序列与一个已知结构的兼容性，如果一个预测所需要的蛋白质折叠不存在于该折叠文库，该方法将失效。

相应程序有：3D-PSSM, Fugue。

➤ Ab Initio Prediction

蛋白质一级序列中包含一些信息可以指导蛋白质找到其天然构象。早期生物物理学研究显示多数蛋白质可自发折叠成处于最低能量附近的稳定结构。这种结构状态叫做蛋白质的天然状态。预测程序使用能量最小化原理。这种算法搜索每一种可能构象来确定一个具有最低全局能量的构象。正如同源建模中侧链精修一样，该方法并不可行。目前的从头预测法将片段搜索法与Threading相结合产生一个结构模型。

相应程序有Rosetta。

总结

	Homology Modeling	Threading	Ab Initio Prediction
原理	一级序列具有较高性的两个蛋白，很可能具有非常相似的三维结构。	蛋白质的结构比其一级序列更保守，故许多不存在序列相似性的蛋白可能具有相似的蛋白折叠	蛋白质的一级序列信息可以指导其正确折叠，天然状态的蛋白质接近其能量的最低状态
结构已知的模板	一级序列相似性很高的模板	高级结构（如二级结构）相似性很高的模板	不需要模板
能量最低化原理	用来优化构建的模型	用来优化构建的模型	预测的主要原理
准确性	最准确，可以提供精修后的原子模型	较准确，只能提供大致的拓扑结构	准确性较差，目前对蛋白质的一级序列与三维结构的关系认识不是很清楚

References

- Bates, P.A., Luo, J. and Sternberg, M.J.E. (1992)
FEBS Lett.301, 207-214
- Jin Xiong. Essential Bioinformatics



Thank you !
