(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101501694 B
(45)授权公告日 2011.11.30

(21)申请号 200780029032.2

(22)申请日 2007.07.24

(30)优先权数据
60/821,553 2006.08.04 US

(85)PCT申请进入国家阶段日
2009.02.03

(86)PCT申请的申请数据
PCT/GB2007/002789 2007.07.24

(87)PCT申请的公布数据
WO2008/015384 EN 2008.02.07

(73)专利权人 英国龙沙生物医药股份有限公司
地址 英国伯克郡

(72)发明人 卡伊·J·科尔霍夫
赫苏斯·苏尔多
米凯莱·文德鲁斯科洛

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 王萍 李春晖

(51)Int.Cl.
C12Q 1/00(2006.01)

(56)对比文件
WO 20050045442 A1,2005.05.19,
PAWAR A P.Prediction of "Aggregation-prone"and "Aggregation-susceptible"Regions in Proteins Associated with Neurodegenerative Diseases.《JOURNAL OF MOLECULAR BIOLOGY》.2005,第350卷(第2期),
DE GROOT NATALIA SANCHEZ.Prediction of "hot sports" of aggregation in disease-linked polypeptides.《BMC STRUCTURAL BIOLOGY》.2005,第5卷(第1期),

审查员 赵婧

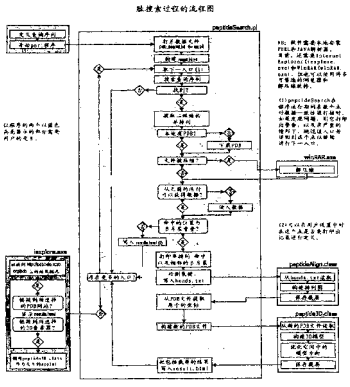
权利要求书 4 页 说明书 17 页 附图 7 页

(54)发明名称

用于对蛋白质聚集进行预测以及设计聚集抑制剂的方法

(57)摘要

本发明提供了一种用于对蛋白质聚集进行预测以及设计聚集抑制剂的方法。这种用于对潜在的蛋白质聚集抑制肽序列进行预测的方法包括步骤:a)识别形成目标蛋白质中的聚集区域的至少一部分的肽序列;b)测定所述肽序列是否形成β-片层的一部分;c)如果步骤b)中获得肯定的结果,则提取该片层的相邻的股;d)识别与所述肽序列相邻的股中的、侧链与所述肽序列相互作用的残基,那些残基形成潜在的蛋白质聚集抑制肽序列。本发明还提供了用于使用以上方法中所识别的残基来设计化合物的方法;这些方法所产生的化合物以及用于实行以上方法的计算机程序。



CN 101501694 B

1. 一种用于对潜在的蛋白质聚集抑制肽序列进行预测的方法，包括步骤：

a) 识别形成目标蛋白质中的聚集区域的至少一部分的肽序列；

b) 在复数个蛋白质结构中测定所述肽序列是否形成 β - 片层的一部分；

c) 如果步骤 b) 中获得肯定的结果，则提取该片层的相邻的股；

d) 识别与所述肽序列相邻的股中的、侧链与所述肽序列相互作用的残基，那些残基形成潜在的蛋白质聚集抑制肽序列。

2. 根据权利要求 1 所述的方法，该方法中，对所述目标蛋白质的复数个外源蛋白质实行所述测定步骤。

3. 根据权利要求 2 所述的方法，其中，使用蛋白质结构的数据库来实行所述测定步骤。

4. 根据权利要求 3 所述的方法，其中，所述测定步骤包括子步骤：识别所述数据库中所包含的、包含与所述肽序列相关的相关肽序列的一组蛋白质；以及在所述组内识别其中所述相关肽序列形成 β - 片层的一部分的那些蛋白质。

5. 根据权利要求 4 所述的方法，其中，所述相关肽序列包括正序的所述肽序列和包括反序的所述肽序列两者。

6. 根据权利要求 4 或权利要求 5 所述的方法，其中，所述相关肽序列包括所述肽序列以及所述肽序列的片段。

7. 根据权利要求 4 所述的方法，其中，所述相关肽序列包括包含有所述肽序列内的一个或更多个氨基酸的保守替换的序列。

8. 根据权利要求 7 所述的方法，其中，以在 PH 7.0 时聚集倾向在彼此 0.2 以内的氨基酸为基础来选择所述保守替换。

9. 根据权利要求 4 所述的方法，其中，所述识别子步骤包括把所述相关肽序列与所讨论的蛋白质的 PDB 文件中的“SHEET”行中所包含的残基相比较。

10. 根据权利要求 4 所述的方法，其中，所述识别子步骤包括识别所述相关肽序列中的、彼此之间形成氢键的那些残基。

11. 根据权利要求 10 所述的方法，其中，为了识别彼此之间形成氢键的那些残基，对至少相距三个残基的每对残基之间的欧几里德距离进行计算，以及如果该距离小于 3.075 埃则假定形成氢键。

12. 根据权利要求 11 所述的方法，其中，使用来自所讨论的蛋白质的 PDB 文件的“ATOM”入口来计算欧几里德距离。

13. 根据权利要求 11 或权利要求 12 所述的方法，进一步包括对所识别的形成氢键的残基对以及它们之间的氢键进行显示的步骤。

14. 根据权利要求 9 或权利要求 10 所述的方法，其中，把所获取的结果与通过实行根据权利要求 11 至 13 中任一权利要求的方法所获取的结果相比较来对所识别的残基进行交叉核对。

15. 根据权利要求 1 所述的方法，其中，步骤 d) 中所识别的残基是那些侧链经由氢键与所述肽序列相互作用的残基。

16. 根据权利要求 1 所述的方法，其中，所述识别步骤使用聚集倾向分布。

17. 根据权利要求 16 所述的方法，该方法中，所述识别步骤在所述聚集倾向分布中选择聚集倾向大于 1 的肽残基。

18. 根据权利要求 1 所述的方法，其中，以实验的方式来执行所述识别步骤。

19. 根据权利要求 1 所述的方法，进一步包括对步骤 d) 中所识别的残基进行显示的步骤。

20. 根据权利要求 19 所述的方法，其中，所述显示步骤包括对所识别的 β－片层中的残基的 3 维排列进行显示。

21. 根据权利要求 1 所述的方法，进一步包括步骤：

测定步骤 d) 中所识别的残基是否与一个或更多个其它的蛋白质相互作用。

22. 根据权利要求 21 所述的方法，该方法中，对所述目标蛋白质的复数个外源蛋白质实行所述测定步骤。

23. 根据权利要求 22 所述的方法，其中，使用蛋白质结构的数据库来实行所述测定步骤。

24. 根据权利要求 23 所述的方法，其中，所述蛋白质结构是介导基本细胞过程的结构。

25. 根据权利要求 23 或权利要求 24 所述的方法，其中，所述测定步骤包括子步骤：识别所述数据库中所包含的、包含与所述肽序列相关的相关肽序列的一组蛋白质；以及在所述组内识别其中所述相关肽序列与所述所识别的残基相互作用的那些蛋白质。

26. 根据权利要求 1 所述的方法，其中，步骤 a) 中所识别的聚集区域的部分是螺旋、环、β－转角或 β－凸起的一部分。

27. 根据权利要求 1 所述的方法，包括产生包括步骤 d) 中所识别的残基的蛋白质聚集抑制肽。

28. 根据权利要求 1 所述的方法，进一步包括步骤：

e) 合成肽库，所述肽库的成员包括步骤 d) 中所识别的残基，以及

f) 确定所述库的成员与目标蛋白质的亲合性。

29. 一种用于产生蛋白质聚集抑制肽的方法，包括步骤：

a) 识别形成目标蛋白质中的聚集区域的至少一部分的肽序列；

b) 测定所述肽序列是否形成 β－片层的一部分；

c) 如果步骤 b) 中获得肯定的结果，则提取该片层的相邻的股；

d) 识别与所述肽序列相邻的股中的、侧链与所述肽序列相互作用的残基，那些残基形成潜在的蛋白质聚集抑制肽序列，

e) 合成肽库，所述肽库的成员包括步骤 d) 中所识别的残基，以及

f) 确定所述库的成员与目标蛋白质的亲合性。

30. 根据权利要求 28 或权利要求 29 所述的方法，包括从所述库把显示出相对于对照物而与目标蛋白质有高亲合性的肽识别为蛋白质聚集抑制肽。

31. 根据权利要求 30 所述的方法，包括对从所述库所识别的肽进行分离。

32. 根据权利要求 30 所述的方法，包括对从所述库所识别的肽进行合成。

33. 根据权利要求 27、权利要求 31 或权利要求 32 所述的方法，包括在蛋白质错误折叠疾病的模型中确定所述肽或肽库的功效。

34. 根据权利要求 33 所述的方法，其中，所述模型包括过表达易聚集蛋白质的细胞。

35. 根据权利要求 34 所述的方法，其中，易聚集蛋白质选自如下的组：野生型 α－突触核蛋白或任何与帕金森症相关联的突变 α－突触核蛋白、亨廷顿蛋白以及其它具有延长的

多聚谷氨酰胺或多聚丙氨酸重复的蛋白质、β 淀粉样蛋白 (Aβ42)、Prion 蛋白、胰岛淀粉样多肽 (hIAPP)、超氧化物歧化酶、Tau 蛋白、α-1- 抗胰蛋白酶以及其它的丝氨酸蛋白酶抑制剂、溶菌酶、玻璃粘连蛋白、晶体蛋白、纤维蛋白原 α 链、载脂蛋白 AI、胱蛋白酶抑制剂、凝溶胶蛋白、乳铁蛋白、角膜上皮蛋白、降钙素、心钠素、催乳素、角蛋白、Medin 或全长乳凝集素、免疫球蛋白轻链、转甲状腺素蛋白 (TTR)、血清淀粉样蛋白 A(SAA)、β2- 微球蛋白、免疫球蛋白重链、或者任何其它与任何蛋白质错误折叠紊乱相关联的蛋白质。

36. 根据权利要求 27、权利要求 31 或权利要求 32 所述的方法，包括确定所述肽执行以下的一个或更多个的能力：

i. 稳定蛋白质以对抗聚集；

ii 减小所存储的蛋白质失去活性的速率；

iii. 降低蛋白质的聚集介导免疫原性；

iv. 增加离体转译系统中蛋白质的产出；

v. 增加用于治疗使用的配方的溶液的稳定性；

vi. 抑制一个或更多个细胞过程；

vii. 防止蛋白质的寡聚或多聚。

37. 一种设计用于蛋白质错误折叠疾病的治疗的聚集抑制剂的方法，包括步骤：根据权利要求 1 至 22 中任一权利要求对蛋白质聚集抑制肽进行预测；以及使用该预测的步骤 d) 中所识别的残基来设计聚集抑制剂。

38. 一种设计用于稳定蛋白质以对抗聚集的化合物的方法，包括步骤：根据权利要求 1 至 22 中任一权利要求对蛋白质聚集抑制肽进行预测；以及使用该预测的步骤 d) 中所识别的残基来设计用于稳定蛋白质以对抗聚集的化合物。

39. 一种用于设计化合物以增加蛋白质的保质期的方法，包括步骤：根据权利要求 1 至 22 中任一权利要求对蛋白质聚集抑制肽进行预测；以及使用该预测的步骤 d) 中所识别的残基来设计化合物以增加蛋白质的保质期。

40. 一种用于设计化合物以减少蛋白质的聚集介导免疫原性的方法，包括步骤：根据权利要求 1 至 22 中任一权利要求对蛋白质聚集抑制肽进行预测；以及使用该预测的步骤 d) 中所识别的残基来设计化合物以减少蛋白质的聚集介导免疫原性。

41. 一种用于设计化合物以增加离体转译系统中蛋白质的产出的方法，包括步骤：根据权利要求 1 至 22 中任一权利要求对蛋白质聚集抑制肽进行预测；以及使用该预测的步骤 d) 中所识别的残基来设计化合物以增加离体转译系统中蛋白质的产出。

42. 一种设计用于使用于治疗使用的配方的溶液稳定性增加的化合物的方法，包括步骤：根据权利要求 1 至 22 中任一权利要求对蛋白质聚集抑制肽进行预测；以及使用该预测的步骤 d) 中所识别的残基来设计具有增加的溶液稳定性的化合物。

43. 一种用于确定化合物对细胞过程的影响的方法，包括步骤：根据权利要求 1 至 22 中任一权利要求对蛋白质聚集抑制肽进行预测；以及针对介导所述细胞过程的蛋白质序列的集合来对该预测的步骤 d) 中所识别的残基进行筛选。

44. 一种用于设计会被用来防止蛋白质的寡聚或多聚的化合物的方法，其中该寡聚或多聚由易聚集区域来介导，包括步骤：根据权利要求 1 至 22 中任一权利要求对蛋白质聚集抑制肽进行预测；以及使用该预测的步骤 d) 中所识别的残基来设计通过与所述易聚集区

域相互作用将会抑制该寡聚或多聚的化合物。

45. 一种用于设计会被用来抑制目标肽或多肽的活性的化合物的方法，包括步骤：

a) 识别形成目标蛋白质中的活性区域的至少一部分的肽序列；

b) 在复数个蛋白质结构中测定所述肽序列是否形成 β - 片层的一部分；

c) 如果步骤 b) 中获得肯定的结果，则提取该片层的相邻的股；

d) 识别与所述肽序列相邻的股中的、侧链与所述肽序列相互作用的残基，那些残基形成潜在的蛋白质聚集抑制肽序列；以及

e) 使用步骤 d) 中所识别的残基来设计会抑制目标蛋白质的活性的化合物。

46. 根据权利要求 37 所设计的化合物，其包括 L- 氨基酸、包括 D- 氨基酸，或者包括 L- 氨基酸和 D- 氨基酸的混合。

47. 根据权利要求 38 所设计的化合物，其包括 L- 氨基酸、包括 D- 氨基酸，或者包括 L- 氨基酸和 D- 氨基酸的混合。

48. 根据权利要求 39 所设计的化合物，其包括 L- 氨基酸、包括 D- 氨基酸，或者包括 L- 氨基酸和 D- 氨基酸的混合。

49. 根据权利要求 40 所设计的化合物，其包括 L- 氨基酸、包括 D- 氨基酸，或者包括 L- 氨基酸和 D- 氨基酸的混合。

50. 根据权利要求 41 所设计的化合物，其包括 L- 氨基酸、包括 D- 氨基酸，或者包括 L- 氨基酸和 D- 氨基酸的混合。

51. 根据权利要求 42 所设计的化合物，其包括 L- 氨基酸、包括 D- 氨基酸，或者包括 L- 氨基酸和 D- 氨基酸的混合。

52. 根据权利要求 43 所设计的化合物，其包括 L- 氨基酸、包括 D- 氨基酸，或者包括 L- 氨基酸和 D- 氨基酸的混合。

53. 根据权利要求 44 所设计的化合物，其包括 L- 氨基酸、包括 D- 氨基酸，或者包括 L- 氨基酸和 D- 氨基酸的混合。

54. 根据权利要求 45 所设计的化合物，其包括 L- 氨基酸、包括 D- 氨基酸，或者包括 L- 氨基酸和 D- 氨基酸的混合。

55. 一种被配置成执行权利要求 1 至 45 中任一权利要求的方法的计算机，其包括

a) 配置成识别形成目标蛋白质中的聚集区域的至少一部分的肽序列的装置；

b) 配置成在多种蛋白质结构中测定所述肽序列是否形成 β - 片层的一部分的装置；

c) 配置成如果步骤 b) 中获得肯定的结果，则提取该片层的相邻的股的装置；

d) 配置成识别与所述肽序列相邻的股中的、侧链与所述肽序列相互作用的残基的装置，所述残基形成潜在的蛋白质聚集抑制肽序列。

56. 根据权利要求 55 的计算机，其中，该计算机被配置成对外部数据库进行访问，所述外部数据库包含获取预测所要使用的已知蛋白质的信息。

用于对蛋白质聚集进行预测以及设计聚集抑制剂的方法

[0001] 本发明涉及用于对蛋白质聚集进行预测以及设计聚集抑制剂的方法。其特别地但并非排他地涉及辅助进行用于稳定蛋白质以对抗聚集的化合物的设计的方法，从而潜在地增加蛋白质的保质期、降低蛋白质的免疫原性以及增加离体转译系统中的产出。

背景技术

[0002] 细胞中或细胞内空间中错误折叠蛋白质的沉积被发现与医学上许多严重的紊乱有关，其中的这些疾病诸如老年痴呆症、帕金森病和 2 型糖尿病。全世界用于治疗那些医学情况的健康服务系统的花费是巨大的，对那些受感染的人们以及对他们的家庭的影响也是巨大的。

[0003] 病例的数量很可能随预期寿命的增长而平稳增加。为解决这一增长中的问题，基于在早期对蛋白质形成聚集的能力进行干扰而开发着新的疗法。

[0004] 细胞中蛋白质的通常的生命周期开始于在核糖体的多肽的合成并继续从初始地、无折叠的状态经由可能包含一个或几个折叠中间体的折叠路径到达蛋白质生物活性的自然态。对于大多数蛋白质，此自然态对应于紧密的折叠构象，虽然是存在一些例外的，其中之一是天然无折叠的 α - 突触核蛋白 (Uversky VN(2002)Natively unfolded proteins : Apoint wherebiology waits for physics, Protein Sci. 11 :739-756)。生命周期结束于变性和降解。

[0005] 细胞对协助进行蛋白质折叠过程的复杂的质量控制机制进行处理。这些机制中的第一个首先是核糖体本身。第二，蛋白质在作为催化剂或助催化剂的热休克蛋白和伴娘蛋白的支持下以正确的方式对蛋白质进行折叠，或者对错误折叠的蛋白质重新进行折叠 (Evans MS, Vlarke TF IV, ClarkPL(2005)Conformations of Co-Translational Folding Intermediates, Prot, Pept. Let. 12(2) :189-195)。

[0006] 在重新折叠失败的情况下，由泛素 - 蛋白酶体系统对错误折叠的蛋白 质进行处理。第一步，把泛素连接到故障结构。这些标签对多肽链进行标记以用于降解，此任务由蛋白酶体来完成。可以在 Dobson CM(2003)Protein folding and misfolding, Nature 426 : 884-890 以及 Vendruscolo M, Zurdo J, Macphee CE, Dobson CM(2003)Protein folding and misfolding :a paradigm of self-assembly and regulation in complex biological systems, Phil. Trans. R. Soc. Lond. A 361 :1205-1222) 中找到对折叠过程和错误折叠过程的更详细的描述。

[0007] 然而，细胞的质量控制会由于各种原因而失败，导致错误折叠的蛋白质的积聚。这些蛋白质随后能够聚集而形成被称为淀粉样纤维、核心区域包括连续的 β - 片层组合的密集结构 (Dobson CM(2005)Prying into prions, Nature 435 :747-749)。

[0008] 在活的组织中，蛋白质沉积（常常是以淀粉样聚集的形式）时常与各种疾病相关联，其中的许多疾病是与年龄有关的。例如，这些疾病包括诸如帕金森病、老年痴呆症和海绵状脑病的神经退化性疾病，以及系统性的（如免疫球蛋白轻链或淀粉样转甲状腺素蛋白）和周围组织的紊乱（如 2 型糖尿病）。在人类中，超过 30 个不同的错误折叠被获知是

与蛋白质沉积相关联的。

[0009] 特别是在预期寿命继续稳步增长的发达的世界中，感染那些疾病的人数的持续增长对社会引起了空前的且日益严重的问题。

[0010] 据估计，2000 年仅美国就约 450 万人感染了老年痴呆症，且病例的数量到 2050 年可能增长到 1600 万 (Hebert LE, Scherr PA, Bienias JL, Bennett DA, Evans DA(2003) Alzheimer Disease in the U.S.Population ;Prevalence Estimates Using the 2000 Census, Arch.Neurol.60 :1119-1122)。人们感染此神经退化性疾病的危险据估计对于超过 60 步年龄的人高达 10 个中 1 个，而对于超过 85 岁的则几乎 2 个中 1 个 (Evans DA, Funkenstein HH, Albert MS, Scherr PA, Cook NR, Chown MJ, HebertLE, Hennekens CH, Taylor JO(1989)Prevalence of Alzheimer's Diseasein a Community Population of Older Persons.Higher than PreviouslyReported, Jama 262 :2551-2556)。对健康系统的影响是非常大的，且一些作者预测神经退化疾病会成为死亡的主要原因 (Lozano AM, Kalia SK(2005)New Movements in Parkinsons's, Sci.Am., 291(1) :58-65)。

[0011] 此外，生物分子在溶液中形成聚集的倾向始终是药品设计中的关键问 题之一。治疗分子不仅必须是可溶性的还必须是反应性的且当以相对高的浓度来进行管理或长时间存放时不应形成聚集。在许多情形下，证明了找到这些多肽足够稳定的条件是耗费时间的且昂贵的，且有时候采用当前现有的方法甚至是不可能的。因此，找到对折叠过程进行干扰的方式以阻止聚集的形成可以改进药品开发的效率。

发明内容

[0012] 因此，为了优选地协助对上述问题进行解决，所期望的是，能够设计与病理蛋白质或者与溶液中的治疗分子相互作用的化合物，从而化合物竞争性地结合到并阻断衍生聚集过程的最重要的部位。对获得此目标的一个逼近将涉及对会干扰聚集过程的肽衍生分子的设计。

[0013] 相应地，最广泛地来说，本发明一方面提供了一种用于设计蛋白质聚集抑制肽的方法，其涉及对侧链会与目标蛋白质中的易聚集区域相互作用的肽序列的识别。

[0014] 本发明的第一方面提供了一种用于对潜在的蛋白质聚集抑制肽序列进行预测的方法，包括步骤 :

[0015] a) 识别形成目标蛋白质中的聚集区域的至少一部分的肽序列 ;

[0016] b) 测定所述肽序列是否形成 β - 片层 (beta-sheet) 的一部分 ;

[0017] c) 如果步骤 b) 中获得肯定的结果，则提取该片层的相邻的股 ;

[0018] d) 识别与所述肽序列相邻的股中的、侧链与所述肽序列相互作用的残基 (residue)，那些残基形成潜在的蛋白质聚集抑制肽序列。

[0019] 优选地，对目标蛋白质的复数个外源蛋白质实行所述测定步骤。相应地，即使目标蛋白质中的聚集区域不形成 β - 片层的一部分，也能够从在其它蛋白质中所找到的相似的或相同的序列以识别合适的肽序列。

[0020] 可以通过使用蛋白质结构的数据库来执行此测定，以及优选地，对所述肽序列是否形成该数据库中所出现的已知蛋白质结构中的任何蛋白质结构中的 β - 片层的一部分进行测定。在本发明的实施例中，使用来自包含大量实验结构和理论模型的结构生物信息

学合作研究协会 (RCSB) 的蛋白质数据银行 (PDB) 的数据 (Berman HM, Westbrook J, Feng Z, GillilandG, Bhat TN, Weissig H, Shindyalov IN, Bourne PE(2000)The Protein Data Bank, Nuc. Acids Res. 28 :235-242)。2005 年 6 月 27 日的时候,PDB 中有 31639 个结构。然而,也可以使用其它的结构数据库或数据银行。

[0021] 对聚集抑制肽进行预测的第一步是识别目标蛋白质中的一个或更多个聚集区域,以及形成此区域的至少一部分的肽序列。

[0022] 用于识别此区域的优选的方法是使用淀粉样聚集分布。DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M(2004)Predictionof the Absolute Aggregation Rates of Amyloidogenic Polypeptide Chains, J. Mol. Biol. 341 :1317-1326 中以及 Pawar AP, DuBay KF, Zurdo J, ChitiF, Vendruscolo M, Dobson CM(2005)Prediction of "aggregation-prone"and"aggregation-susceptible"regions in proteins associated withneurodegenerative diseases, J. Mol. Biol. 350 :379-392 中对用于对多肽链内的聚集热点区域进行预测的这一理论方法进行了描述。该方法提供了能够被用来基于氨基酸的许多固有属性来为任意蛋白质计算淀粉样聚集倾向分布的算法。

[0023] 一旦获取了目标蛋白质的淀粉样聚集分布,通过对该分布中聚集倾向超过预定值 (如 1) 的蛋白质部分进行考虑就能够识别聚集区域。

[0024] 可替选地或额外地,可以通过实验测量来识别聚集区域,例如,通过肽或蛋白质或其片段中每个残基的系统突变,并把这些肽或蛋白质或片段的片段进行合成并以离体试验对它们的聚集倾向进行分析。

[0025] 优选地,所述测定步骤包括子步骤 :识别蛋白质结构数据库中所包含的、包含与形成目标蛋白质中的聚集区域的至少一部分的肽序列相关的相关肽序列的一组蛋白质 ;以及在所述组内识别其中所述相关肽序列形成 β - 片层的一部分的那些蛋白质。

[0026] 为了增加可能的命中 (hit)(即,所识别的其相关肽序列形成 β - 片层的一部分的蛋白质) 的数量,相关肽序列优选地包括所关注的肽序列以及所述肽序列的片段。

[0027] 可替选地或额外地,为增加可能的命中的数量,相关肽序列可以包括包括具有所关注的肽序列的一个或更多个氨基酸的保守替换的序列。

[0028] 在此步骤的上下文中,保守替换是保留了被替换氨基酸的聚集属性的替换。具体地,可以以在 PH 7.0 时聚集倾向在彼此 0.2 以内的氨基酸为基础来选择保守替换。可替选地或额外地,可以以具有相似属性 (例如, 碱性、酸性、疏水性、极性、芳香性) 的残基为基础来选择保守替换。

[0029] 更优选地,相关肽序列既包括正序的也包括反序的一个或更多个上述的肽序列。

[0030] 本方面的一个实施例中,识别数据库中所包含的一组蛋白质的子步骤包括把相关肽序列与所讨论的蛋白质的 PDB 文件中的"SHEET"(片层) 行中所包含的残基相比较。

[0031] 可替选地或额外地,识别数据库中所包含的一组蛋白质的子步骤包括识别所述相关肽序列中的、彼此之间形成氢键 (hydrogen bonds) 的那些残基。

[0032] 一个可以识别那些彼此之间形成氢键的残基的方法是,对至少相距三个残基的每对残基之间的欧几里德距离进行计算,如果该距离小于 3.2 埃,以及更优选如果该距离小于 3.075 埃,则假定形成氢键。在一个实施例中,使用来自所讨论的蛋白质的 PDB 文件的"ATOM"入口 (entry) 来计算欧几里德距离。

[0033] 优选地，该方法包括对所识别的彼此之间形成氢键的残基对以及它们之间的氢键进行显示的进一步的步骤。

[0034] 为了核对来自以上用于识别蛋白质组的方式之一的结果的有效性，来自两个方法的结果都可以用来对从其它方法所识别的残基进行交叉核对。可替选地或额外地，可以用其它用于识别蛋白质组的方法来执行这些交叉核对。

[0035] 该方法优选地包括对步骤 d) 中所识别的残基进行显示的进一步的步骤，以及还可以包括对关于这些被识别的残基所来自的蛋白质的信息进行显示的步骤。此显示可以采用所识别的 β - 片层中的残基的 3 维排列的形式。

[0036] 优选地，该方法包括进一步的步骤 :对不直接参与与所述形成目标蛋白质中的聚集区域的一部分的肽序列的相互作用的、与潜在的蛋白质聚集抑制肽序列相邻的侧链和股的主干进行修饰，以使与所述肽序列的相互作用最大化以及增加潜在的蛋白质聚集抑制序列的潜在的聚集抑制属性。

[0037] 一旦识别了一个或更多个肽序列或“模板”，则可以设计以及合成肽库，其在不直接涉及与所讨论的聚集区域的相互作用的模板区域引入了变异性。随后可以使用专用的生化聚集和细胞毒性试验来对此库进行筛选， 以对出现各种化合物的情况下蛋白质的毒性以及聚集率的变化进行研究。

[0038] 优选地，通过给候选氨基酸序列添加对诸如稳定性和可溶性的属性进行改进的修饰来构建库。

[0039] 例如，本文中所描述的方法可以包括 :e) 合成肽库，所述肽库的成员包括步骤 d) 中所识别的残基，以及 f) 确定所述库的成员与目标蛋白质的结合亲合性。

[0040] 在库内可以识别一个或更多个相对于对照物以高亲合性结合到目标蛋白质的肽。这些肽会是候选的蛋白质聚集抑制肽。

[0041] 可以把被发现以高亲合性结合到目标的肽分离、纯化和 / 或合成。

[0042] 可以以对细胞过程的干扰对上述所预测的 (predicted) 或所识别的蛋白质序列进行筛选 (即，毒理学)。例如，一种方法可以包括 :测定以上步骤 d) 中所识别的残基是否与蛋白质数据银行 (或任何其它的蛋白质数据库) 中存在的一个或更多个非目标蛋白质相互作用，优选地是能够介导诸如代谢途径、离子稳态结构蛋白、涉及对应激的反应的蛋白质、调控基因表达、DNA 修复等的基本 (essential) 细胞过程的非目标蛋白质。

[0043] 可以对所述目标蛋白质的复数个外源蛋白质实行所述测定步骤，优选地使用蛋白质结构的数据库。例如，识别数据库内的、包含与测定肽序列相关的相关肽序列的一组蛋白质 ;并在该组内识别其中相关肽序列与以上步骤 d) 中所识别的残基相互作用的那些蛋白质。

[0044] 候选的蛋白质聚集抑制肽序列可以被识别为不与介导基本细胞过程的蛋白质相互作用。

[0045] 可以在蛋白质错误折叠疾病的模型中确定上述所识别的肽序列的功效。

[0046] 蛋白质错误折叠疾病的模型是本领域中公知的。适合的模型包括过表达易聚集蛋白质的细胞、以及过表达易聚集蛋白质的诸如老鼠或果蝇模型的转基因动物模型以及那种可能会也可能不会面临诸如氧化应激等其它挑战的模型。

[0047] 例如，见 :

[0048] Junn E, Mouradian MM, Human alpha-synuclein over-expressionincreases intracellular reactive oxygen species levels and susceptibility to dopamine, Neurosci Lett. 2002 ;320 :146-50。

[0049] Lev N, Melamed E, Offen D, Proteasomal inhibition hypersensitizesdifferentiated neuroblastoma cells to oxidative damage,Neurosci Lett. 2006 ;399 :27-32。

[0050] McGowan E, Eriksen J, Hutton M, Adecade of modeling Alzheimer' sdisease in transgenic mice, Trends Genet. 2006 ;22 :281-9。

[0051] Whitworth AJ,Wes PD,Pallanck LJ,Drosophila models pioneer anew approach to drug discovery for Parkinson's disease, Drug Discov Today. 2006 年 2 月 ; 11(3-4) :119-26。

[0052] 易聚集蛋白质包括如下的蛋白质、前体或片段 :α - 突触核蛋白 (野生型或任何与帕金森症相关联的突变)、亨廷顿蛋白 (以及其它具有延长的多聚谷氨酰胺或多聚丙氨酸重复的蛋白质)、β 淀粉样蛋白 (Aβ42)、Prion 蛋白、胰岛淀粉样多肽 (hIAPP)、超氧化物歧化酶、Tau 蛋白、α -1- 抗胰蛋白酶以及其它的丝氨酸蛋白酶抑制剂、溶菌酶、玻璃粘连蛋白、晶体蛋白、纤维蛋白原 α 链、载脂蛋白 AI、胱蛋白酶抑制剂、凝溶胶蛋白、乳铁蛋白、角膜上皮蛋白、降钙素、心钠素、催乳素、角蛋白、Medin(或全长乳凝集素)、免疫球蛋白轻链、转甲状腺素蛋白 (TTR)、血清淀粉样蛋白 A(SAA)、β 2- 微球蛋白、免疫球蛋白重链、或者其它任何与任何蛋白质错误折叠紊乱相关联的蛋白质。

[0053] 可以对上述所识别的肽序列执行以下的一个或更多个的能力进行确定 :

[0054] i. 稳定蛋白质以对抗聚集 ;

[0055] ii 减小所存储的蛋白质失去活性的速率 ;

[0056] iii. 降低蛋白质的聚集介导免疫原性 ;

[0057] iv. 增加离体转译系统中蛋白质的产出 ;

[0058] v. 增加用于治疗使用的配方的溶液的稳定性 ;

[0059] vi. 抑制一个或更多个细胞过程 ;

[0060] vii. 防止蛋白质的寡聚或多聚。

[0061] 本发明进一步的方面提供了用于使用以上第一方面的步骤 d) 中所识别的残基来设计聚集抑制剂以治疗蛋白质错误折叠疾病的方法 ;用于使用 所识别的残基来设计用于稳定蛋白质以对抗配方、溶剂和其它溶液中的聚集的化合物 (例如,生物药品、抗体、酶等) 的方法 ;用于使用所识别的残基来设计化合物以增加这些蛋白质的保质期的方法 ;用于使用所识别的残基来设计化合物以降低蛋白质由于聚集的免疫原性的方法 ;以及用于使用所识别的残基来设计化合物以增加离体转译系统中蛋白质的产出的方法。

[0062] 本发明的另一方面提供了计算机程序,该计算机程序在计算机上运行时执行以上方面中任一方面的方法。

[0063] 本发明的另一方面提供了包含根据之前方面的计算机程序的计算机数据载体。

[0064] 本发明的进一步的方面提供了被配置成执行以上方法方面中任一方面的方法的计算机。优选地,此计算机被配置成对数据库进行访问,这些数据库包含获取预测所要使用的已知蛋白的信息。可以把这些数据库存储在本地,例如,存储在硬盘驱动器上或存储在存

储器中，但优选地对其进行远程存储并通过诸如网络或因特网的通信链接来进行访问。

附图说明

[0065] 现在将参考附图对本发明的实施例进行描述，其中：

[0066] 图 1 示出了 Aβ42 和 α-突触核蛋白在 PH 7 的淀粉样聚集分布；

[0067] 图 2 示出了 γ-晶体蛋白 D 的突变在 PH 7 的淀粉样聚集分布；

[0068] 图 3 示出了 mutCRYD34-58 的残基 1 至 8 的 PeptideSearch（肽搜索）程序的输出；

[0069] 图 4 示出了肽搜索过程的流程图；

[0070] 图 5 示出了从 PDB 获取的 pdb seqres.txt 以及 ss.txt 文件的节选；

[0071] 图 6 示出了 β-片层中三个相邻的股中的氢键的预测；

[0072] 图 7 示出了用于对残基之间的氢键进行预测的程序的实例输出；

[0073] 图 8 示出了肽搜索中命中的结果概况；以及

[0074] 图 9 示出了对找到 (found)PDB 中的短肽匹配的位置 (position) 进行了标识的 mutCRYD34-58、Aβ42 和 α-突触核蛋白的聚集倾向分布。

具体实施方式

[0075] 下面将对本发明的实施例进行描述。将用 γ-晶体蛋白 D(mutCRYD) 的突变来证明本实施例的方法论。为了示意的目的，还对两个进一步的蛋白质进行了论述：涉及帕金森病的 α-突触核蛋白，以及与老年痴呆症相关联的 Aβ42。

[0076] 人类眼睛的晶状体细胞中有丰富的 mutCRYD。当错误折叠时，它能够形成呈现为白内障的聚集，导致视力模糊或失明 (Héon E, Priston M, Schorderet DF, Billingsley GD, Girard PO, Lubsen N, Munier FL(1999)The γ-Crystallins and Human Cataracts：A Puzzle Made Clearer, Am. J. Hum. Genet. 65：1261-1267，以及 Dahm R(2004)Dying to see, Sci. Am., 291(4)：52-59)。γ-晶体蛋白 D 与该突变有三个残基不同，R58H、R36S 以及 R14C，它们都被发现会增加聚集 (Pande A, Pande J, Asherie N, Lomakin A, Ogun O, King J, Benedek GB(2001)Crystal cataracts：Human genetic cataract caused by protein crystallization, PNAS, 98(11)：6116-6120)。

[0077] α-突触核蛋白是 140 个残基的蛋白质，其被发现是路易体（帕金森病人的大脑中所发现的会引起神经退化的密集的沉积 -Spillantini MG, Schmidt ML, Lee VMY, Trojanowski JQ, Jakes R, Goedert M(1997) α-Synuclein in Lewy bodies, Nature. 388：839-840) 中的主要成分。

[0078] Aβ42 是 42 个残基的小疏水肽，在老年痴呆症的突触功能中断这一神经病状最常见的形式中起直接作用 (Selkoe DJ(2002)Alzheimer's diseaseis a synaptic failure, Science 298：789-790)。

[0079] 形成密集的 β-片层状结构的倾向，即聚集的倾向，在整个蛋白质中有所不同，取决于氨基酸组成和序列。

[0080] 本实施例方法的目的是识别候选的肽序列，这些候选的肽序列允许通过对目标蛋白质的易聚集区域进行封锁 (block) 来减少聚集形成的速度和量。

[0081] 聚集区域的识别

[0082] Pawar AP, DuBay KF, Zurdo J, Chiti F, Vendruscolo M, Dohson CM(2005) Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases, J.Mol.Biol.350 :379-392 对 α - 突触核蛋白和 Aβ42 这两个蛋白质部分中的聚集的所谓的"敏感区域"进行了预测。

[0083] 更一般地,以上论文提出了允许对蛋白质的聚集倾向分布进行计算的算法,且该算法显示出了所给出的结果与广泛范围的实验测量是相当吻合的。对于任何给定的肽序列,它允许对各单个残基的倾向或者对平滑地通过数个氨基酸的窗口的倾向进行计算。表 1 分别列出了每个氨基酸在 PH7.0 的倾向。

[0084] 表 1 :20 个天然氨基酸中每个氨基酸分别在 PH 7.0 的聚集倾向。

[0085]

残基	倾向		残基	倾向
Trp	2.92		Gly	-3.96
Phe	2.80		His	-4.31
Cys	1.61		Ser	-5.08
Tyr	1.03		Gln	-6.00
Ile	0.93		Asn	-6.02
Val	0.49		Asp	-9.42
Leu	-0.25		Lys	-9.55
Met	-1.06		Glu	-10.38
Thr	-2.12		Arg	-11.93
Ala	-3.31		Pro	-11.96

[0086] 图 1 示出了根据此算法所计算的 α - 突触核蛋白和 Aβ42 的聚集倾向分布(以上的论文中也给出了)。左边的图示出了 Aβ42 在 PH 7 的聚集倾向分布。未作改变的野生型 Aβ42 序列的聚集分布被绘制为中间的线,同时还有对沿该序列每个氨基酸位置的任何可能的突变进行测定后所获得的最大的和最小的倾向值。画出了 $Z^{prof}_{agg} = 1$ 的线来辅助对促聚集区域进行识别。通过实验过程所生成的同样的区域被示为阴影区域。

[0087] 类似地,右边的图示出了 α - 突触核蛋白在 PH 7 的聚集倾向分布,也包括 $Z^{prof}_{agg} = 1$ 的线来辅助对促聚集区域进行识别。在此图中,蛋白质被认为是纤维结构的大的区域由浅灰色阴影示出 ;高度淀粉样 NAC 区域由更深的灰色阴影示出 ;被发现是该 NAC 区域内显著的淀粉样段的 69-79 区域由散列的线条示出。

[0088] 数值穿过阈值(例如,图 1 中的 $Z^{prof}_{agg} = 1$)的蛋白质部分被认为是具有最高的聚集倾向并被假定是在体纤维形成和聚集的核心区域。根据图 1 所预测的核心聚集区域在此实施例中被用作用于对任何可以作为抑制剂的相互作用对象进行检测的目标序列。

[0089] 将用 mutCRYD 来证明本实施例的方法论。

[0090] 图 2 示出了 mutCRYD 的聚集倾向分布,取七个残基的平均。如图 1 中的聚集倾向分布一样,添加了聚集倾向值 1.0 处的线来辅助在被认为对聚集敏感的区域与蛋白质的其它部分之间进行区分。

[0091] 为了进一步的搜索,选择了残基 34 至 58。此区域具有预测聚集热点区域的两个顶点并包含两个突变。为了完整的分析,对所有聚集倾向大于 1 的区域都会进行考虑,但所选择的短片段已足以对所使用的方法进行证明。24- 残基的查询序列是 SARVDSGCWMLYEQPNYSGLQYFL(SEQID NO :1),将被称为 mutCRYD34-58。

[0092] 搜索 - 概述

[0093] 虽然对单个文件进行搜索来搜索序列是标准的流程且能够用大多数的文本编辑程序或命令行工具来实行，但文件中可能会容易出现上百个或上千个短序列的命中。手工地对它们中的每个进行考虑、寻找二级结构 (structure) 信息、以及随后核对 PDB 网站以求更多信息、以及使用 3D 工具来显现结构以及寻找残基间的相互作用以提取经由氢键紧密链接的肽，对于每个蛋白质会花费数星期或数月的时间。

[0094] 因此，开发了专用软件来对该过程进行处理。该软件被开发为大量的小工具，随后逐步被合并到称为 PeptideSearch 的单个的、综合的程序中。

[0095] 选择了编程语言 PERL 用于编写 PeptideSearch，因为所有所使用的数据都来自纯文本文件且 PERL 提供了对那些文件进行解析以及提取和处理其中所包含的信息的强大的例程。PERL 还是免费可用的软件，由于这些原因而在生物信息学者中流行。

[0096] 图 4 呈现了本实施例的方法论以及软件的流程图，并提供了对下面的方法的详细阐述的概述。

[0097] PeptideSearch 靠调用许多的外部程序来扩展其功能性。图 4 中也示出了这些调用。

[0098] 在程序的使用中，用户的输入是目标序列加上许多参数，该程序最终生成包括肽和它们的相互作用对象的三维可视图的候选的肽的简要概况。因此，识别任意蛋白质可能的抑制剂所需的时间从潜在的几个星期缩短到几个小时。

[0099] PeptideSearch 构建被称为 'result.html' 的 HTML 文件来保存每个命中的记录 (record)。由程序和外部工具所构建的此文件以及所有其它的文件被存储在新的子目录中。此目录的名字来源于当前的日期和时间，从而磁盘上保留有所有程序的运行结果以用于将来的使用。在执行搜索时，PeptideSearch 还把命中的有关信息打印到命令提示或控制台，同时还有额外的状态消息。一旦程序结束其运行，则 'result.html' 包含对所有所找到的命中的概览，使得对候选的肽进行选择是容易的。图 3 中呈现了对 mutCRYD34-58 的残基 1 至 8 进行搜索后程序的输出。

[0100] 搜索 - 详细的描述

[0101] 对于本实施例，搜索是以结构生物信息学合作研究协会 (RCSB) 的蛋白质数据银行 (PDB) 中的入口为基础的，因为这一可免费访问的在线的源包含大量的实验结构和理论模型。因此，在 PBD 中的一个肽序列内找到查询序列 (query sequence) 的全部或一部分给出了对相关结构的访问，允许对二级结构和相互作用对象进行识别。

[0102] 搜索的第一阶段是在 PDB 的所有的蛋白质入口中来搜索目标序列。这可以通过以下方式达到 :在本地对 PDB 文件进行镜像并分别对它们中的每个进行解析以求残基的序列。然而，此途径在时间和空间方面都是没有效率的 :2005 年 6 月 27 日的时候，PDB 中有 31639 个结构，其中的许多是具有数个链的多聚体。

[0103] 此实施例所采用的替选方式是采用包含所有序列信息的单个文件一次并随后对其进行搜索以检测匹配。从 RCSB 的 PDB ftp 服务器可获取 该文件 :

[0104] ftp://ftp.rcsb.org/pub/pdb/derived_data。

[0105] pdb_seqres.txt 文件是 FASTA 格式的所有 PDB 序列的列表。类似地，在同样的地址可以找到包含 FASTA 格式的 PDB 中所有二级结构的文件 ss.txt。本实施例所使用的文件

版本的日期是 2005 年 6 月 27 日。

[0106] 二级结构文件呈现了初步的且快速的源，以执行序列到二级结构的匹配以及对 β - 片层构象中是否涉及命中进行检测。

[0107] 图 5 是来自两个文件的简短的节选 :分别来自 seq_res. txt 和 ss. txt 的前十行。下面的表 2 示出了 ss. txt 中所使用的缩写的含义。

[0108] 表 2

[0109]

缩写	结构元素		缩写	结构元素
H	螺旋		T	氢键转角
E	伸展 β 股		S	弯曲
G	310 螺旋		B	被分离的 β 桥中的残基
I	π 螺旋			

[0110] 对序列文件和结构文件进行清理

[0111] 以上所提及的两个文件是由 PDB 通过使用来自所有的 PDB 入口的数据而自动衍生出来的。由许多不同的程序确定了 PDB 文件中的二级结构信息。

[0112] 然而，两个文件的长度是不相等的。图 5 中的摘选示出了其中的一个原因 :序列文件中的 FASTA 标题行更长，包含更详细的信息。

[0113] 考虑到数据的来源，期望对每个序列有一个结构入口且反之亦然。此外，首先，似乎文件中的行数可以是相同的，使得对两个文件中的内容进行匹配很方便。

[0114] 然而，这些假设都不是真的 :行数并非总是匹配的。在文件上运行的、提取所有 FASTA 标题行的 PERL 脚本对于 pdb_seqres. txt 返回 74560 个入口，对于 ss. txt 返回 69903 个入口。在把文件用作输入之前，还有许多其它的问题需要解决。例如，ID 的使用在两个文件内以及在两个文件之间是不一致的。序列文件使用诸如 1tsv_ 的 ID，并以 1thl_、1thl_1、1thl_2，或以 1pr2_A、1pr2_B 来区分不同的链，而 ss. txt 则使用 1TSV :_、1THL :_、1THL :_ :1、1PR2 :A、1PR2 :B。此外，序列文件中的许多 ID 在二级结构文件中没有相应的入口，且反之亦然 :诸如 1J3W、1OG8 和 1VGS 的入口包含在 ss. txt 中但不在 pdb-seqres. txt 中。它们由蛋白质数据银行中的 ID 2CVZ、2BUH 和 2CV4 所取代，其仅考虑了 pdb_seqres. txt。

[0115] 间是不一致的。序列文件使用诸如 1tsv_ 的 ID，并以 1thl_、1thl_1、1thl_2，或以 1pr2_A、1pr2_B 来区分不同的链，而 ss. txt 则使用 1TSV :_、1THL :_、1THL :_ :1、1PR2 :A、1PR2 :B。此外，序列文件中的许多 ID 在二级结构文件中没有相应的入口，且反之亦然 :诸如 1J3W、1OG8 和 1VGS 的入口包含在 ss. txt 中但不在 pdb-seqres. txt 中。它们由蛋白质数据银行中的 ID 2CVZ、2BUH 和 2CV4 所取代，其仅考虑了 pdb_seqres. txt。

[0116] 把所有的唯一入口从两个文件移除要采用许多 PERL 脚本。然后，作为结果的文件包含 68383 个匹配的入口。

[0117] 由于一些序列在长度上不匹配，潜在的问题仍然存在。PeptideSearch 程序被配置成在运行期间指出这些不一致。

[0118] 把查询匹配到序列

[0119] 所关注的实验结构是那些其中出现了来自查询串的氨基酸序列的结构。这意味着，需要对正向的串的精确的匹配进行搜索，对于 mutCRYD34-58 ：

[0120] SARVDSGCWMLYEQPNYSGLQYFL(SEQ ID NO :1)

[0121] 以及对于逆向的序列，此处 ：

[0122] LFYQLGSYNPQEYLMWCGSDVRAS(SEQ ID NO :2)。

[0123] 较之为短肽找到精确的序列匹配，在 PDB 中为长肽找到精确的匹配明显是更加不可能的。实际上，当为整个蛋白质搜索匹配时，很可能仅会找到一个匹配 ：该蛋白质自身。

[0124] 即使当集中关注于蛋白质的敏感区域时，也会想要搜索出具有那些只是太长而不能在 PDB 中给出精确命中的氨基酸的延伸 (stretch) 的命中。由于 PDB 中不包括 γ - 晶体蛋白 D 的突变的结构，所以对于完整的 24- 残基的序列没有结果。

[0125] 为增加命中的产出，实施了两个进一步的方法。

[0126] 第一，目标蛋白质序列被分成子串，为这些子串中的每个子串搜索相互作用对象。这允许把结合到序列的一部分的肽识别出来。随后可以对这些分别进行测定，例如，与其它的肽相结合，或者通过把几个这样的肽连接在一起造成更长的肽。

[0127] 在 PeptideSearch 程序中，用户可以设置搜索序列的最小长度 n 以及最大长度 m。随着试验，在找到仍然足够长而仍然有用的相互作用对象时，发现了长度范围在 5 至 8 个残基的查询序列给出了充足的命中数量。然而，也可以使用其它长度的子串。

[0128] 程序被配置成自动对长度在给定范围中的初始序列的所有邻接子串运行完整的搜索。因此，为长度 l 的完整的查询序列所搜索的查询序列的总数量 q 是 ：

[0129] $$q = \frac{(l-n+1)\cdot(l-n+2)-(l-m)\cdot(l-m+1)}{2}$$

[0130] 第二，程序允许使用正则表达式。这允许把变异性引入搜索序列。相应地，PeptideSearch 定义保守替换，当对氨基酸序列文件进行搜索时对这些保守替换进行考虑。

[0131] 例如，根据表 1，Gln(Q) 和 Asn(N) 具有相似的聚集倾向，因此，用以增加有用命中的数量的一个选择将会是 ：在搜索中对两个都进行考虑。例如搜索串 QANT(SEQ ID NO :3) 的正则表达式将会是 [QN]A[QN]T(SEQ ID NO :3-6)。可以以类似的方式来使用其它的基于残基的聚集属性的替换。

[0132] 两个方法可以合并，从而把每个下至最小长度的可能的子串转换成考虑上述保守替换的正则表达式。

[0133] 找到二级结构元素

[0134] 来自 pdb_seqres.txt 和 ss.txt 的数据都是在程序开始时被读取一次并被存储在大的数组中。随后对来自两个数组的入口联合进行处理。如以上所注意到的，由于氨基酸序列和结构的长度并非总是匹配，所以行数不能被用作引导。而是，用 FASTA 标题行作为导向，以确保搜索对共同所属的入口进行考虑以及确保两个文件中的信息保持同步。

[0135] 如果入口被分在几行，则通过把它们之间的换行符去掉来对它们进行联接。这为每个入口产生了两个串变量，一个用于完整的氨基酸序列以及一个用于各自的二级结构。在这点上，如果两个串的长度不等，则不再认为给定的信息是可靠的。在此情况下，程序显示警告来使用户意识到不一致并建议对命中进行人工检查。

[0136] PERL 提供了用于对序列串进行搜索以对查询序列的出现进行搜索的简单易用的例程，要么是寻找完美的匹配、要么是通过对正则表达式进行匹配。如果有任何的命中存在，则 PeptideSearch 把它们从序列串中提取出来并用二级结构串的相应区域对它们进行排列。这允许首先对匹配中有多少残基是在伸展的 β - 片层构造中的进行目视的核对。

[0137] 由于结构的附近 (neighbourhood) 也会是所关注的，该程序还被配置成输出匹配的左边和右边的许多结构信息和残基。此窗口的尺寸可以在程序选项中来进行设置。

[0138] 理想地，此信息可以仅仅被用来对所有那些其二级结构文件指示存在 β - 片层的入口进行选择且随后仅对那些进行进一步的研究。然而，用所使用的文件，这会导致许多误报及许多漏报。其主要原因是 ss.txt 中的预测时常不准确。当把排列与 PDB 文件中的实际结构相比较时会发现这一点。

[0139] 因此，可见最好对序列中的每个命中都进行考虑，而不论 ss.txt 中的信息。这增加了所必须处理的命中的数量，进一步需要访问 PDB，但它也确保了没有重要的候选肽被遗漏。

[0140] 获取 PDB 文件

[0141] 一旦在一个入口中检测到查询序列，则需要获取更多的信息且必须要对实际的 PDB 文件进行处理。从 FASTA 标题行来提取 PDB ID。PeptideSearch 中定义了本地的 PDB 目录。

[0142] 如果本地目录中有属于此 ID 的 PDB 文件，压缩的或未压缩的，则使用此本地的拷贝。否则，PeptideSearch 使用其例程中的一个例程来自动访问剑桥晶体学数据中心 (CCBC) 的 RCSB 蛋白质数据银行 ftp 服务器的镜像，ftp://pdb.ccdc.cam.ac.uk/rcsb/data/structures/divided/pdb/，并下载 PDB 文件的压缩拷贝。随后经由到程序 WinRAR 的系统调用来对此 .Z-archive 进行解压。

[0143] 处理 PDB 文件

[0144] 在打开 PDB 文件之前，程序对之前的命中是否使用了同样的 PDB ID 进行核对。如果同样的氨基酸序列内有数个命中，或者当文件中的结构是多聚体从而每个链上有相似的命中存在时，会是这种情况。如果 ID 是同样的，则数据仍包含在程序的数据数组内并能够被重新使用，节省了大量的时间。否则，打开 PDB 文件的本地拷贝，把数据读取到存储器。

[0145] 每个文件有两个部分是所关注的。第一是能够被用来对查询序列与 β - 片层的重叠进行识别的行。这些行以示出 β - 片层位置的“SHEET”标识符开始。

[0146] 第二指的是给出了分子中每个颗粒的坐标的“ATOM”入口。此信息允许对氢键进行预测，因为知道哪些残基实际上相互作用是提供了重要的信息的并在设计抑制剂肽上给了研究者更多的自由。例如，侧链不涉及相互作用的残基会被其它的为肽提供更好的生化属性的所取代。

[0147] β - 片层中的每个股都有一个“SHEET”行。这包含股中第一个和最后一个残基的索引。如果此股在片层中不是第一个，则该行还包含把此股和前面的股进行注册 (register) 的两个残基，即，被发现被氢键连接的单个的一对残基的索引。在 http://www.rcsb.org/pdb/docs/format/pdbguide2.2/guide2.2_frame.html 可以找到有关“SHEET”和“ATOM”入口的完整信息。

[0148] 从那些行提取信息，氨基酸序列中的命中的起始点和末尾点提供了对重叠的测

定：如果片层的起始索引不大于命中的末尾的索引且片层的末尾索引不小于命中的开始，则存在重叠，在该情况下，对重叠的残基的数量进行计算。

[0149] 理想地，重叠的残基的数量不应太小，因为它限定了潜在的抑制剂肽的长度。仅包括少量残基的抑制剂不会很有效率，且设计出具有良好属性的稳定的肽将是困难的。程序使得用户可以定义阈值并把所有的重叠短于此阈值的命中从‘result.html’的输出中排除。

[0150] 随后，在考虑了股的方向（平行或反平行）和注册的排列中，对与查询序列相重叠的 β 股以及相邻的股（一个或两个，取决于第一个股在片层中的位置）进行显示。

[0151] 例如，如果肽序列是‘ADDYYTATGHWYAT”(SEQ ID NO:7) 以及股分别经过 (run) 残基 1 至 4、6 至 10、以及 12 至 14，是在 3 和 7、以及 9 和 13 具有注册的反平行 β-片层构造，则排列的结果是：

[0152] ADDY(SEQ ID NO:8)

[0153] HGTAT(SEQ ID NO:9)（反向）

[0154] YAT

[0155] 在‘result.html’中，在此排列中的查询序列的残基随后会被示为红色。比方说，如果要寻找出在以上的肽序列中的索引 7 至 10 可以找到的‘ATGHW’(SEQ ID NO:10)，则残基‘HGTA’(SEQ ID NO:11) 在该排列中会被高亮显示。

[0156] 排列允许迅速对那些在查询序列与 β-片层之间具有大的重叠的命中进行筛选。它示出了有多少以及有哪些残基形成 β-片层，以及因此，如果它们的相互作用对象要被结合到短的抑制剂肽中的话，有多少以及有哪些残基会被封锁。

[0157] 虽然以上的搜索方法学提供了识别候选肽的一个途径，但对于当前的可用数据它不是十分完善的。具体地，‘SHEET’行中所找到的残基的索引引用‘ATOM’线中的索引，而不引用同样的文件中的‘SEQRES’记录。蛋白质数据银行保证了对‘ATOM’记录的引用进行测定以求正确性。然而，‘pdb_seqres.txt’中的数据衍生自‘SEQRES’记录，且它们的索引常常不一样。例如，‘SEQRES’入口以 MET 开始，而第一个‘ATOM’入口是 SER，或者序列是同样的，但第一个‘ATOM’入口以残基索引 21 开始，而不是 1。

[0158] 这意味着，虽然以上示出排列的技术可以正确地重新产生相对的 β-片层的股位置，但残基的标识可能取自氨基酸序列的错误的部分。这还意味着，对与 β-片层的重叠的识别可能首先就不正确：查询序列实际上可以在结构中位于不同的位置。

[0159] peptideAlign.class

[0160] 为确保当随后查看此命中的概况时能够发现该事件，使用‘ATOM’入口开发了进一步的排列方法。这允许用可替选的方法来衍生候选的肽，或者用于对从上述‘SHEET’搜索所获取的结果进行交叉核对。

[0161] 此排列方法还包括对残基之间的氢键进行检测和显示，使得该排列更加具有信息性。

[0162] 设计出了如下算法用于在‘ATOM’数据中进行搜索：对于至少相距三个残基并在所关注的 β-片层部分内的所有原子对，对它们之间的欧几里德距离进行计算（使用标准公式$d=\sqrt{\Delta x^2+\Delta y^2+\Delta z^2}$，其中 d 是欧几里德距离，Δx、Δy 和 Δz 是各坐标的差）。

[0163] 如果所计算的距离小于 3.075 埃，则该对残基的索引被添加 (push) 到氢键存储数

组。截止长度选自文献，且是足够大的以对主干原子与侧链之间的大多数氢键进行检测，而且是足够小的以保持低的漏报数量。

[0164] 完成预测之后，股和键信息被写入被称为‘bond.txt’的文件。此文件被用作中介的数据存储用以与另一应用共享信息：被构建来用于画出图形化的 β－片层排列的基于Java的工具‘peptideAlign.class’。虽然PERL具有强大的例程来处理大量的文本，但Java提供了更容易使用的图形化功能。因此，通过使‘peptideSearch.pl’使用系统调用来链接到PeptideAligh而把两个编程语言的属性进行了合并。后者构建排列的截屏，PERL程序随后把其加入到‘result.html’中的当前命中的概览中。

[0165] 使用图形方便了对残基之间的氢键进行标识，这可以通过简单地画线来做到。PeptideAlign不直接关注反平行的 β－片层构造，但如果像是在氢键穿过彼此的情形中那样应当对股的方向进行反转，则图形输出是进行了标识的。例如，图6中标识氢键的蓝色的线示出了需要反转中间的股的方向。进一步的算法被用来使股的方向正确并把它们相对于彼此进行移动以使键的总长度最小化。

[0166] PeptideAlign从‘bonds.txt’读取肽和键信息，连同预期输出图像文件的名称。图像文件名称是hitX.png形式的，其中X表示命中的连续数量。程序把文件中的信息翻译成图形化表示并把控制返回到peptideSearch。图7中示出了实例输出：文件包含预计要保存截屏的文件的路径和名称，文件中的原子和键的编号(number)，对于所有原子后面跟着残基ID，以及关于它们中的哪些是由键来键合的信息；右边的图形化表示示出了哪些残基是彼此键合的。

[0167] 最低限(Minimalist)PDB文件

[0168] 由于PDB的处理(curation)，由‘ATOM’搜索所生成的排列是精确的。如果这对应于由PeptideSearch所构建的一个，则‘SEQRES’记录就是一致的。否则，将需要用户用肉眼对图形概览中的重叠进行核对。如果没有，则需要对PDB文件进行人工核对，以查看是否有不同的 β－片层包含了查询序列的一部分或全部。

[0169] 目前，注意到的是，不可能从‘ATOM’记录重新构建氨基酸序列并对此进行搜索以求出现查询串。这是因为记录中常常缺失残基。例如，在pdb1p9w.ent中，残基108后面所跟着的是残基121。可以用许多方式对间断进行解释，例如，由于实验数据不足、NMR或X射线的分辨率低、或者高度灵活的侧链导致信号模糊使得不可能精确确定原子的位置。另一解释仅仅是，提供数据的研究有时候仅关注于蛋白质的特定部分，例如活性部位，并因此没有对剩余的结构提供任何预测。然而，可获得更完备的数据的情况下，这种搜索方法将是可独立应用的、以上所讨论的‘SHEET’搜索的合适的替选。

[0170] 在 β－片层中找到了命中的话，则查看残基中的哪些相互作用以及建立用于设计候选肽的支架是可能的。例如，不与查询序列中的残基相互作用的残基会由其它的氨基酸来取代，允许了要进行设计和测定的肽的范围更大。

[0171] 例如，在图6的短排列中，SARV(SEQ ID NO:12)与VTY相互作用。预测氢键是在A与V，以及V与Y之间。然而，T似乎并不与查询序列相互作用，所以可以把肽VXY用作用于构建候选肽的支架，可以用不同的氨基酸来取代X。实际中，该支架太短而无法起作用，因此优选地搜索长度为5个和更多个残基的肽。

[0172] 对每个成功的命中，PeptideSearch构建新的.pdb文件，其中包含 β－片层中所

涉及的所有原子的坐标。为做到这点，它把完备结构中包含了具有查询序列片段及其相互作用对象的 β 股的那部分提取出来。

[0173] 文件可以用诸如 Rasmol 或 vmd 的任何通常的 3D 查看器来打开，并准许对所预测的侧链的氢键和空间方向有效率地进行验证。

[0174] Peptide3D.class

[0175] 为给出新的 .pdb 文件中结构的预览，截屏被加到‘result.html’的记录中。为做到这样，需要这样的 3D 查看器 :允许反映出 (render) 分子、把它们放进空间中的有信息性的方向中、以及进行截屏，所有这些都不需要用户的交互。

[0176] Results.html

[0177] 一旦 PeptideSearch 通过 pdb_seqres.txt 中的所有入口完成了对所有可能的子串和突变的运行，它就用系统调用把浏览器窗口打开并在概览中呈现出所有它已经找到的命中。此概览为与每个命中相关的 PDB ID 提供 了到 PDB 概况页的链接并提供了到 Peptide3D 的链接，该 Peptide3D 是以 applet 来实施的、允许以 800×600 象素格式对结构进行查看。

[0178] 结果

[0179] 用 mutCRYD34-58 作为输入来运行程序、把最小和最大子串长度设置为五、不允许突变并排除所有与 β - 片层的重叠小于四的命中，在 pdb_seqres.txt 中产生了 153 个精确的匹配，其中的 47 个是在 β - 片层中。

[0180] 这些结果中仅八个结果长于三个残基且后者被加进在‘result.txt’中。事实上，此集合实际上仅给出了两个不同的命中，因为前两个以及最后六个匹配实际上是相同的，只是由于它们来自多聚体的不同的链。

[0181] 那些命中内的第一个是指Putatinve Glycine Cleavage SystemTranscriptional Repressor(PDB ID 1U8S，Putatinve 甘氨酸断裂系统转录阻抑物）中所找到的序列 SARVD(mutCRYD34-58 残基 1 至 5 ;SEQ IDNO :13)。图 3 中呈现出了这一命中的结果 (result)。图 6 中对四残基重叠进行了描绘。

[0182] 在 Endo-1,4-Beta-Xylanase II 中的序列 PNYSG(mutCRYD34-58 残基 15 至 20 ;SEQ ID NO :14) 找到了第二个命中。图 8 中示出了来自 HTML 文件的此命中的结果的概况。

[0183] 图 9 在 mutCRYD34-58 的聚集倾向分布中对两个命中进行了高亮显示。图 9 还在各自的分布中示出了与 α - 突触核蛋白和 Aβ 42 这两个蛋白质的聚集热点区域相邻近或相接近的精确匹配的位置。对于 α - 突触核蛋白，子串的长度被设置为 6 至 7 的范围，以及对于 Aβ 42，它被设置到 5。

[0184] 对于 mutCRYD，第一个命中作为聚集抑制剂的候选不会非常有效，因为它没有位于聚集的敏感区域内。然而，第二个命中具有封锁肽在此部位开始形成聚集的潜力。

[0185] 对于其它的两个蛋白质，命中所对应的相互作用对象中的每个相互作用对象都具有作为聚集抑制剂的潜力。当子串长度被设置为 5 时，α - 突触核蛋白的命中的数量大大增加。

[0186] 对 α - 突触核蛋白进行分析的结果是本申请人进一步专利申请的主题。

[0187] 使用上述软件设计了一系列肽来与 α - 突触核蛋白的区域 61-66(EQVTN ;SEQ ID NO :15) 和 71-76(VTGVT ;SEQ ID NO :16) 相互作用。

[0188] 此分析识别了包括与 α-突触核蛋白的残基 61-66(EQVTN ;SEQ IDNO :15) 之间的区域相互作用的 D- 氨基酸序列的肽。一个适合的肽包括或包含 D- 氨基酸序列 QYSVLI(列表中不含的 D- 氨基酸)(下面的表 3 中所描述的 ZP-0195)。其它适合的肽包括或包含具有一个、两个或三个氨基酸替换的 D- 氨基酸序列 QYSVLI。例如,肽可以包含的 D- 氨基酸序列选自如下的一组,该组包含 :qykvli、qysvpi、qyspli、qypvli、rysvli、qysvli、qytvli、pysvli、或 qysvlv(列表中不含的 D- 氨基酸)。

[0189] 肽可以包括一个、两个或三个额外的 N 端残基。

[0190] 例如,肽可以包括或包含的序列选自如下的一组,该组包含 :ekysvli 和 drysvli。

[0191] 分析还识别了包括与 α-synuclein 的残基 71-76(VTGVT ;SEQ ID NO :16) 之间的区域相互作用的 D- 氨基酸序列的肽。例如,肽可以包含 D- 氨基酸序列 hhviva(ZP-0158) 或者可以包括或包含具有一个、两个或三个氨基酸替换的 D- 氨基酸序列 hhviva。优选地,不对 N 端组氨酸残基进行替换。

[0192] 例如,肽可以包括或包含的序列选自如下的一组,该组包含 :hhvvva、hhvlva、hhvkva、hhveva、hpviva、hhvivp、hhvivv、hhvivt、hhvivy、hhvivw、hhtivv、hhtivk、hhtvva、hhtlva、hhtlvv、hhtevy 以及 hhttvy。

[0193] 对于区域 61-65,肽 AC-qysvli-NH2(ZP-0195) 被设计来相互作用以及防止聚集。对此序列的变异也进行测定,其中在任意给定位置的一个或更多个氨基酸被另一个所替换,变异是在 N- 末端进行的,如添加额外的氨基酸和乙酰 (ZP-0195 至 ZP-0230)。

[0194] 对于区域 71-75,肽 AC-hhviva-NH2(ZP-0158) 被设计来相互作用以及防止聚集。对该序列的变异也进行测定,其中在任意给定位置的一个或更多个氨基酸被另一个所替换。所有被测定的肽都是 N 端乙酰化的且序列开始处的 2 个组氨酸在所有的设计中保持恒定 (ZP-0158 至 ZP-0194)。

[0195] 通过用 50μM ASYN 和 100μM 抑制剂、用 50mM tris 和 150mMNaCl 以及 20μM 琉黄素 T 实行聚集试验来对所有的肽对 TBS 中 ASYN 聚集的抑制进行测定。反应体积是 200μL。用仅具有 ASYN 的和仅具有缓冲剂的对照物 (with ASYN only and buffer only controls) 来在 96 孔的聚丙烯板中建立每个反应。在 37C 摇晃 48 小时来孵化反应,通过读取琉黄素 T 荧光来对聚集进行监控。

[0196] 使用把数据拟合到 Sigmoidal 函数 f(x) = k+A/(1+exp(-b(t-t0))) 的 Zyentiafit 软件对运动轨迹进行拟合,从而可以计算停滞时间、聚集的速率和 ThT 荧光的总的变化。

[0197] 根据肽的有效性给它们分级,下面的表 3 示出了那些为进一步的研究而选择的序列。选择是基于停滞时间的增加大于 20%和 / 或 ThT 荧光或聚集速率的降低大于 20%的肽。

[0198] 表 3- 被设计来与 ASYN 的区域 61-66 以及 71-76 相互作用的肽序列,其示出了离体防止 ASYN 聚集的有效性。

[0199] (列表中不含的 D- 氨基酸)

[0200]

Zyentia 码	序列	ThT 荧光的 降低%	聚集速率的 降低%	停滞期的增 加%
ZP-0158	Ac-hhviva-NH2	15.1	18.4	32.5
ZP-0159	Ac-hhvvva-NH2	19.8	18.6	36.6

ZP-0160	Ac-hhvlva-NH2	8.8	10.5	36.5
ZP-0161	Ac-hhvkva-NH2	-5.9	2.3	19.7
ZP-0162	Ac-hhveva-NH2	6.3	6.6	33.4
ZP-0164	Ac-hpviva-NH2	3.7	-2.3	43.0
ZP-0168	Ac-hhvivp-NH 2	-8.6	-5.4	29.8
ZP-0169	Ac-hhvivv-NH2	-0.5	-13.3	26.1
ZP-0170	Ac-hhvivt-NH2	24.5	29.3	29.4
ZP-0171	Ac-hhvivy-NH2	13.9	13.2	43.6
ZP-0172	Ac-hhvivw-NH2	-1.7	6.9	28.1
ZP-0175	Ac-hhtivv-NH2	21.9	16.8	89.7
ZP-0179	Ac-hhtivk-NH2	0.6	-0.6	30.4
ZP-0180	Ac-hhtvva-NH2	19.8	18.4	30.3
ZP-0181	Ac-hhtlva-NH2	22.8	-1.2	29.4
ZP-0186	Ac-hhtlvv-NH2	30.7	29.7	16.1
ZP-0193	Ac-hhtevy-NH2	8.7	-15.8	31.4

[0201]

Zyentia码	序列	ThT荧光的 降低%	聚集速率的 降低%	停滞期的增 加%
ZP-0194	Ac-hhttvy-NH2	11.6	7.7	38.8
ZP-0202	Ac-qykvli-NH2	53.5	54.5	-28.8
ZP-0204	Ac-qysvpi-NH2	1.4	-5.6	23.6
ZP-0205	Ac-qyspli-NH2	16.6	-11.9	30.6
ZP-0206	Ac-qypvli-NH2	3.3	8.4	25.9
ZP-0207	Ac-qpsvli-NH2	5.5	3.6	29.7
ZP-0212	Ac-rysvli-NH2	51.1	44.9	-14.4
ZP-0213	Ac-ekysvli-NH2	25.6	12.9	29.4
ZP-0214	Ac-drysvli-NH2	20.4	4.8	-12.6
ZP-0215	NH3-qysvli-NH2	27.6	17.5	3.1
ZP-0221	NH3-qytvli-NH2	24.7	22.8	-22.4
ZP-0222	NH3-qykvli-NH2	42.3	34.3	7.7
ZP-0228	NH3-pysvli-NH2	40.8	46.2	38.2
ZP-0229	NH3-qysvlv-NH2	17.8	24.3	33.5

[0202] 根据结果来设计化合物

[0203] 通过上述方法或软件工具所识别的命中可以被用作用来设计聚集抑制剂或稳定剂的模板。可以使用分子动态计算方法或其它适合的计算方法来基于这些模板对化合物的库进行测定以求它们与目标序列的亲合性。

[0204] 计算方法可以把具体的力场和能量极小化例程应用于基于模板的各种抑制剂，以使得抑制剂与目标聚集多肽之间的相互作用最大化。(见 DasB, Meirovitch H, Navon IM, Performance of hybrid methods for large-scaleunconstrained optimization as applied to models of proteins, J Comput. Chem. 2003 ;24 :1222-31 以及 de Bakker PI, Depristo MA, Burke DF, Blundell TL, Ab initio construction of polypeptide fragments :Accuracy ofloop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model, Proteins, 2003 ;51(1) :21-40)。

[0205] 可以用聚集试验离体对所选择的化合物库进行测定来识别会抑制目标多肽聚集的先导 (lead) 化合物。

[0206] 当稳定剂/聚集抑制剂需要抗其它蛋白酶时，可以使用逆对映体 (retro-enantio) 衍生物 (反向 C-N 序列和 D- 氨基酸)。

[0207] 可以把所识别的稳定剂 / 聚集抑制剂与其它的蛋白质 / 肽融合。与稳定剂 / 聚集抑制剂融合的蛋白质 / 肽可以 :作为载体用以针对目标输送到身体的具体位置、具体器官、具体细胞类型等 ;促进细胞内的输送 ;促进透过血脑屏障 ;增加血浆半衰期 ;以及 / 或者与另一蛋白质或受体相互作用。

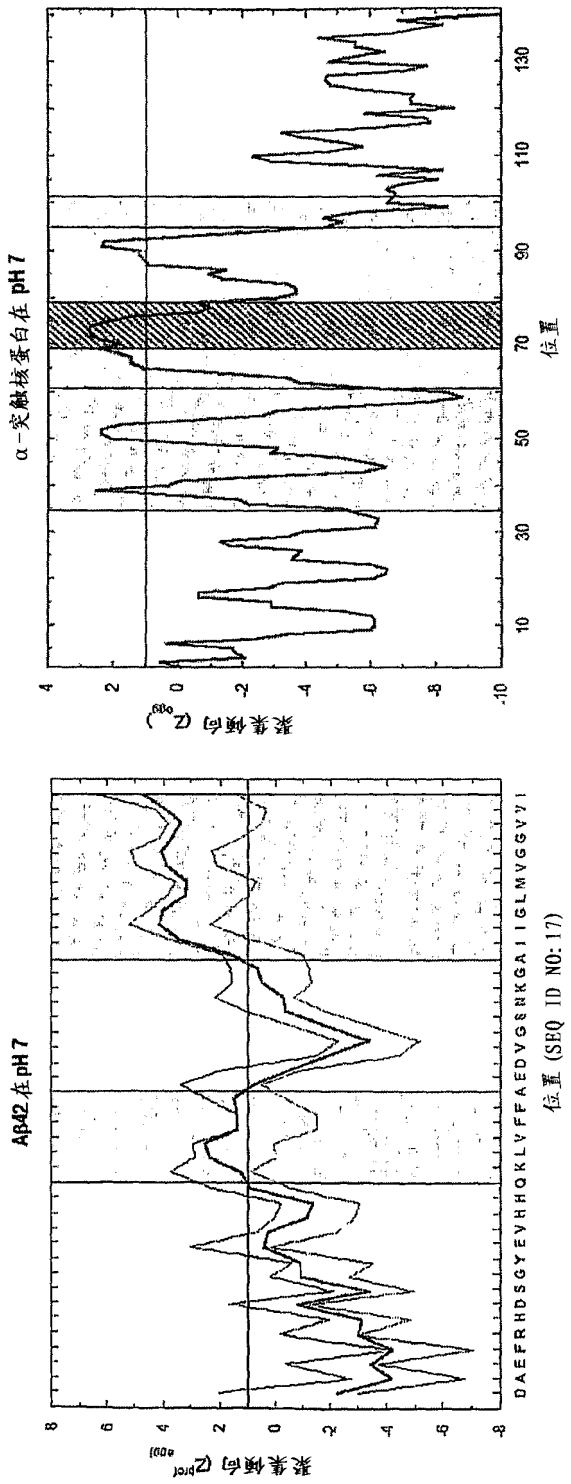

α-突触核蛋白在 pH 7
聚集倾向 (Z_{agg})
位置
Aβ42 在 pH 7
聚集倾向 (Z_{aggl}^{Prof})
DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVI
位置 (SEQ ID NO: 17)

图 1

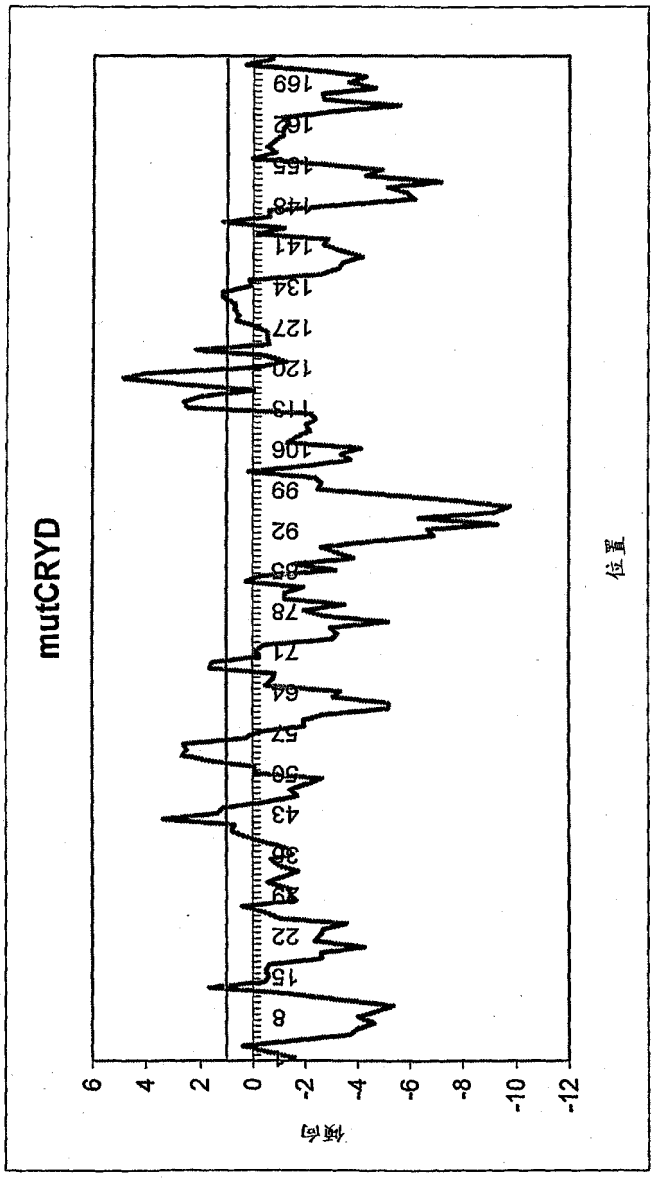
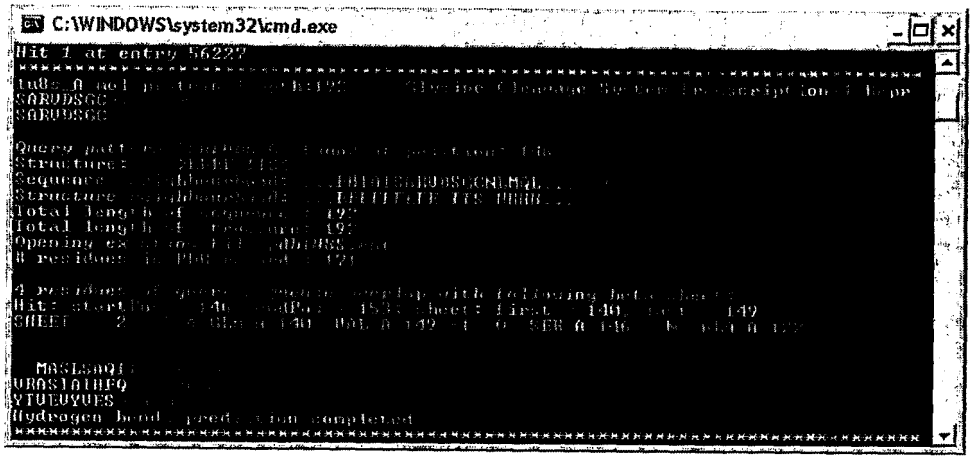

mutCRYD
倾向
6
4
2
0
-2
-4
-6
-8
-10
-12
位置
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
127
134
141
148
155
162
169

图 2

a)


C:\WINDOWS\system32\cmd.exe
Hit 1 at entry 56227
SARVDSGC
SARVDSGC
Hydrogen bond prediction completed

b)

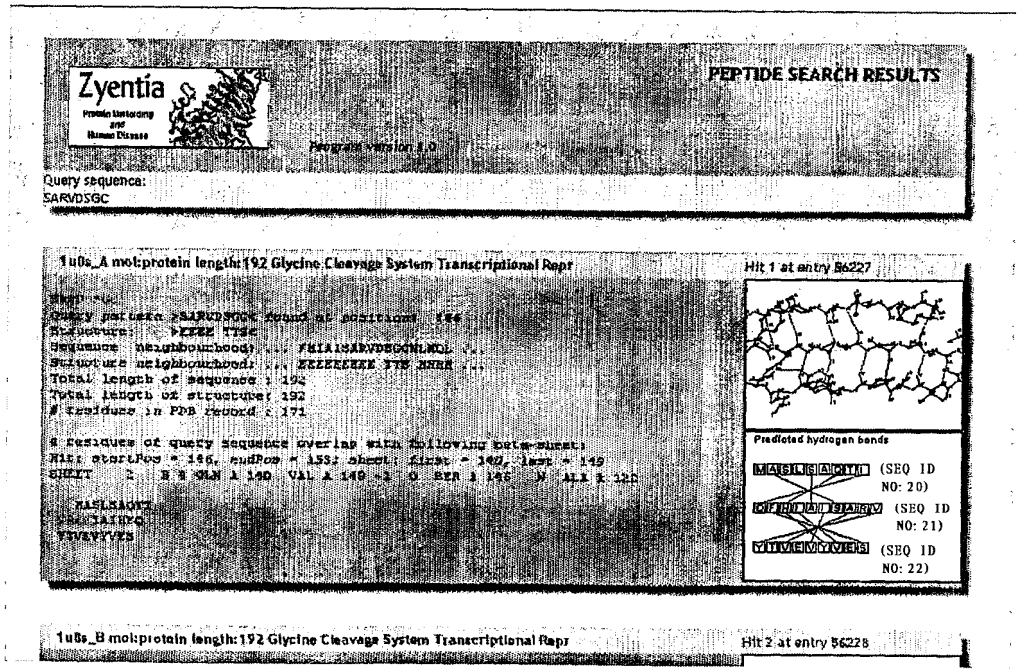

Zyentia
PEPTIDE SEARCH RESULTS
Query sequence:
SARVDSGC
1u8s_A mol:protein length:192 Glycine Cleavage System Transcriptional Repr
Hit 1 at entry 56227
Predicted hydrogen bonds
(SEQ ID NO: 20)
(SEQ ID NO: 21)
(SEQ ID NO: 22)
1u8s_B mol:protein length:192 Glycine Cleavage System Transcriptional Repr
Hit 2 at entry 56228

图 3

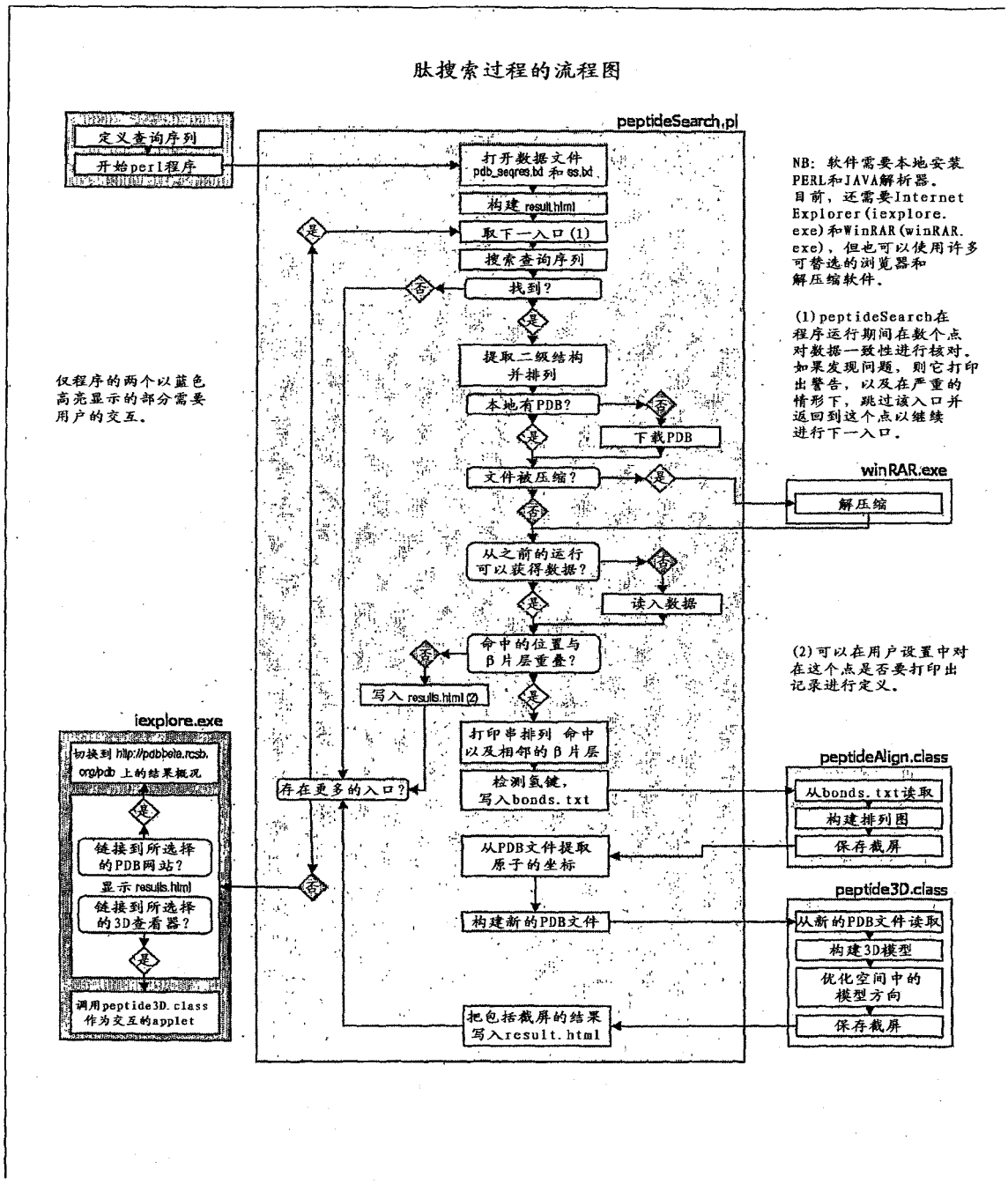

肽搜索过程的流程图
定义查询序列
开始perl程序
peptideSearch.pl
打开数据文件 pdb_seqres.txt 和 ss.txt
构建 result.html
取下一入口 (1)
搜索查询序列
找到?
提取二级结构并排列
本地有PDB?
下载PDB
文件被压缩?
winRAR.exe
解压缩
从之前的运行可以获得数据?
读入数据
命中的位置与β片层重叠?
写入 results.html (2)
打印串排列 命中以及相邻的β片层
检测氢键, 写入bonds.txt
peptideAlign.class
从bonds.txt读取
构建排列图
保存截屏
从PDB文件提取原子的坐标
构建新的PDB文件
peptide3D.class
从新的PDB文件读取
构建3D模型
优化空间中的模型方向
保存截屏
把包括截屏的结果写入result.html
存在更多的入口?
iexplore.exe
切换到 http://pdbbeta.rcsb.org/pdb 上的结果概况
链接到所选择的PDB网站?
显示 results.html
链接到所选择的3D查看器?
调用peptide3D.class作为交互的applet
仅程序的两个以蓝色高亮显示的部分需要用户的交互。
NB: 软件需要本地安装PERL和JAVA解析器。目前，还需要Internet Explorer(iexplore.exe)和WinRAR(winRAR.exe)，但也可以使用许多可替选的浏览器和解压缩软件。
(1)peptideSearch在程序运行期间在数个点对数据一致性进行核对。如果发现问题，则它打印出警告，以及在严重的情形下，跳过该入口并返回到这个点以继续进行下一入口。
(2)可以在用户设置中对在这个点是否要打印出记录进行定义。

图 4

来自pdb_seqres.txt的前十行:

```
>1pr2_A mol:protein length:239     Purine Nucleoside Phosphorylase (SEQ ID NO:23)
MATPHINAEMGDFADVVLMPGDPLRAKYIAETFLEDAREVNNVRGMLGFTGTYKGRKISVMGHGMGIPSCSIYTK
ELITDFGVKKIIRVGSCGAVLPHVKLRDVVIGMGACTDSKVNRIRFKDHDFAAIADFDMVRNAVDAAKALGIDAR
VGNLFSADLFYSPDGEMFDVMEKYGILGVEMEAAGIYGVAAEFGAKALTICTVSDHIRTHEQTTAAERQTTFNDM
IKIALESVLLGDKE
>1pr2_B mol:protein length:239     Purine Nucleoside Phosphorylase (SEQ ID NO:23)
MATPHINAEMGDFADVVLMPGDPLRAKYIAETFLEDAREVNNVRGMLGFTGTYKGRKISVMGHGMGIPSCSIYTK
ELITDFGVKKIIRVGSCGAVLPHVKLRDVVIGMGACTDSKVNRIRFKDHDFAAIADFDMVRNAVDAAKALGIDAR
VGNLFSADLFYSPDGEMFDVMEKYGILGVEMEAAGIYGVAAEFGAKALTICTVSDHIRTHEQTTAAERQTTFNDM
IKIALESVLLGDKE
```

来自ss.txt的前十行:

```
>1PR2:A
  BTTB   TTSS SEEEEESSTHHHHHHHHHT BS EE B GGG  EEEEEETTEEEEEE   SSHHHHHHHHH
HHHHHS   EEEEEEEEEE STTTTTT EEEESEEEES SHHHHHTTTS    B  HHHHHHHHHHHHHTT  EE
EEEEEE S SS S TTHHHHHHHTT  EEESSHHHHHHHHHHTT EEEEEEEE EETTS     HHHHHHHHHH
HHHHHHHHHHH
>1PR2:B
   SS    TTSS SEEEE SSHHHHHHHHHHH BS EE B GGG  EEEEEETTEEEEEE   SSHHHHHHHHH
HHHHHS   EEEEEEEEEE STTTTTT EEEESEEEES SHHHHHTTTS    B  HHHHHHHHHHHHHTT   E
EEEEEE S SS SSTTHHHHHHHTT  EEESSHHHHHHHHHHTT EEEEEEEE EETTT     HHHHHHHHHH
HHHHHHHHHHH
```

图 5

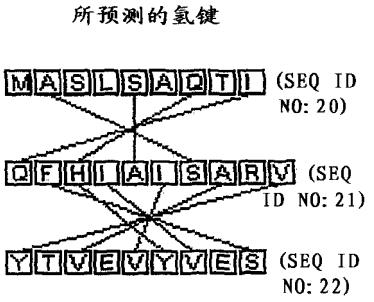


图 6

```
OUTPUTTO 2005_08_19_14h11m01s/images/hit38.png
NUMATOMS 9
NUMBONDS 4
RESID THR 188
RESID VAL 189
RESID SER 190
RESID ILE  60
RESID ASN  61
BOND  ASN  61SER 188
RESID PHE  62
RESID SER  63
BOND  SER  63THR 190
RESID GLY 148
BOND  GLY 148PHE  61
RESID SER 149
BOND  SER 149ASN  62
```

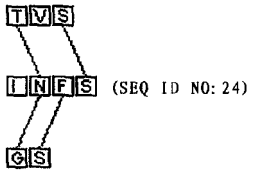


图 7

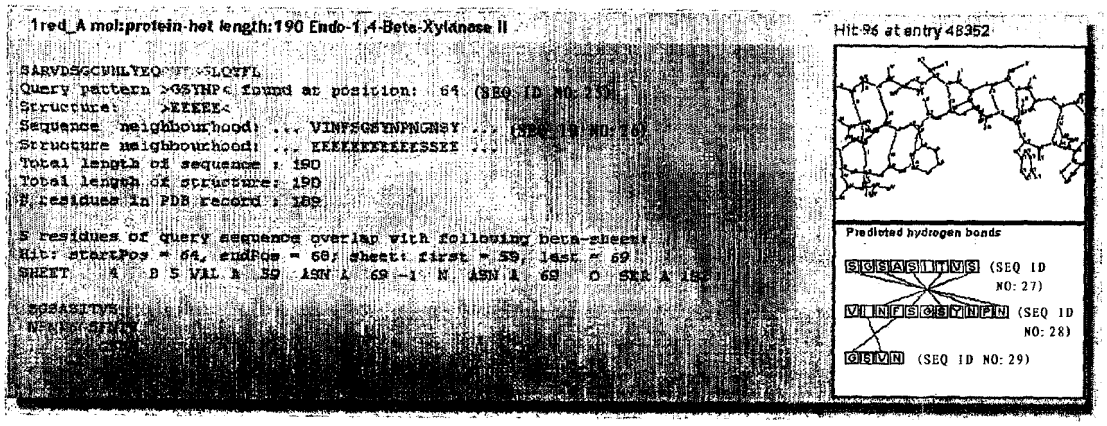

1red_A mol:protein-het length:190 Endo-1,4-Beta-Xylanase II
Hit 96 at entry 48352
Predicted hydrogen bonds
SQSASITVS (SEQ ID NO: 27)
VINFSGSYNPN (SEQ ID NO: 28)
QSVN (SEQ ID NO: 29)

图 8

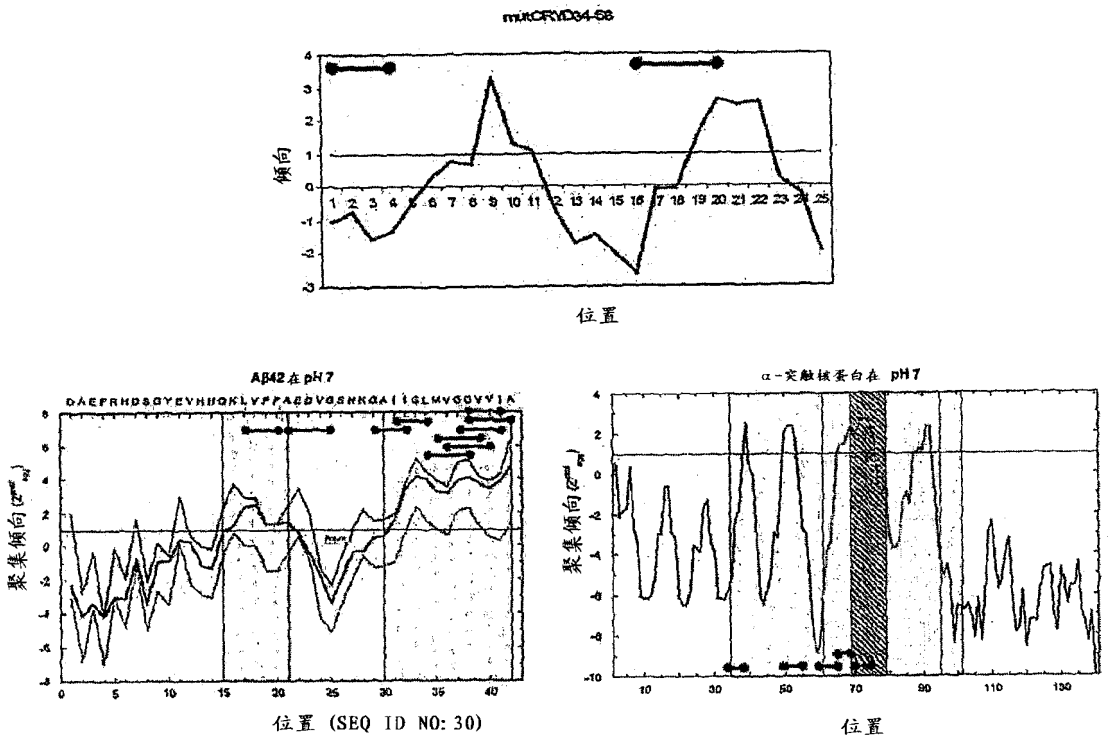

mutCRYD34-58
倾向
位置
Aβ42在pH7
聚集倾向
位置 (SEQ ID NO: 30)
α-突触核蛋白在 pH7
聚集倾向
位置

图 9