

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

C07K 14/00 (2006.01)

C12N 15/00 (2006.01)

C12Q 1/68 (2006.01)



[12] 发明专利说明书

专利号 ZL 02816138.6

[45] 授权公告日 2006 年 12 月 13 日

[11] 授权公告号 CN 1289522C

[22] 申请日 2002.8.17 [21] 申请号 02816138.6

[30] 优先权

[32] 2001. 8.17 [33] US [31] 60/313,402

[32] 2002. 4.22 [33] US [31] 60/374,355

[86] 国际申请 PCT/KR2002/001560 2002.8.17

[87] 国际公布 WO2003/016571 英 2003.2.27

[85] 进入国家阶段日期 2004.2.17

[73] 专利权人 图尔金株式会社

地址 韩国大田广域市

[72] 发明人 金晋秀 裴光熙 朴卿顺 权宁道

柳银铉 黄文宣

审查员 赵彦豪

[74] 专利代理机构 永新专利商标代理有限公司

代理人 林晓红

权利要求书 4 页 说明书 211 页 附图 13 页

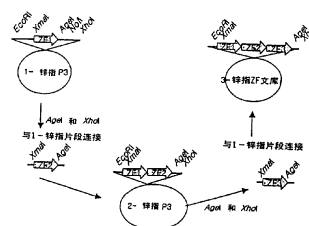
[54] 发明名称

锌指结构域文库

[57] 摘要

本发明揭示了嵌合锌指结构域文库。该文库可包括来自天然存在的蛋白质的两或多个锌指结构域，所述蛋白质例如是哺乳动物蛋白质，尤其是人类蛋白质。从该文库中可以鉴别有用的嵌合锌指结构域。本发明还揭示了识别特定定位点的锌指结构域的氨基酸序列。

HA Tag
GACCAAGCT TGGCACCATG GTGTACGCTT AGGAGGTGGC CGACTAGCGG GAATTGGCTC
HIS
CAAAAAGAA GAGAAAGTA GGCATCCGAA TTCAAGCGGC CGATAGAT CTGAGCATG
XbaI
CATCTAGAGG G



1. 一个包含众多多肽的文库，每个多肽均包含一个第一个和一个第二个锌指结构域，

其中（1）每个多肽的第一个和第二个锌指结构域均与来自一天然存在的蛋白质的一个锌指结构域相同，并且（i）不存在于该同一天然存在的蛋白质中，或者（ii）存在于该同一天然存在的蛋白质中但呈与其在所述多肽中的构象不同的构象，（2）第一个锌指结构域在所述众多多肽之中变化，（3）第二个锌指结构域在所述众多多肽之中变化。

2. 权利要求 1 的文库，其中所述众多多肽中每个多肽均以解离常数（ K_d ）小于 5nM 与靶 DNA 位点结合。

3. 权利要求 1 的文库，其中所述众多多肽中的至少一个多肽的第一个锌指结构域选自表 5、6 和 7 任一所示的锌指结构域。

4. 权利要求 3 的文库，其中所述众多多肽中的每个多肽的第一个锌指结构域选自表 5、6 和 7 任一所示的锌指结构域。

5. 权利要求 3 的文库，其中所述众多多肽中的至少一个多肽的第一个和第二个锌指结构域选自表 5、6 和 7 任一所示的锌指结构域。

6. 权利要求 1 的文库，其中所述天然存在的蛋白质是一种真核细胞蛋白质。

7. 权利要求 6 的文库，其中所述天然存在的蛋白质是一种哺乳动物蛋白质。

8. 权利要求 7 的文库，其中天然存在的蛋白质是一种人蛋白质。

9. 权利要求 1 的文库，其中所述众多多肽中的每个多肽均固定在一个固体支持物上。

10. 权利要求 1 的文库，其中所述众多多肽中的每个多肽均展示在病毒或病毒颗粒表面。

11. 权利要求 1 的文库，其中所述众多多肽中的每个多肽均还包

含一个第三个锌指结构域。

12. 权利要求 11 的文库，其中所述第三个锌指结构域是一天然存在的蛋白质的结构域。

13. 权利要求 11 的文库，其中所述第三个锌指结构域不是一天然存在的蛋白质的结构域。

14. 权利要求 13 的文库，其中所述第三个锌指结构域通过插入、缺失或取代不超过 6 个氨基酸而与一天然存在的蛋白质的结构域不同。

15. 权利要求 1 的文库，其中所述众多多肽中的每个多肽还包含一个转录调节结构域。

16. 权利要求 1 的文库，其中所述众多多肽包含至少 100 个不同多肽。

17. 一个包含众多多核苷酸的文库，每个多核苷酸均编码来自权利要求 1-8 或者 11-16 任一项所述的文库的一个不同多肽。

18. 权利要求 17 的文库，其中每个多核苷酸是一个质粒或噬菌粒的节段。

19. 权利要求 17 的文库，其中每个多核苷酸位于一细胞内。

20. 权利要求 19 的文库，其中所述细胞是一种真核细胞。

21. 权利要求 20 的文库，其中所述细胞是酵母细胞。

22. 权利要求 19 的文库，其中所述细胞含有一种异源报道基因构建体，其包含与一启动子可操纵地连接的报道基因。

23. 权利要求 17 的文库，其中每个多核苷酸包装进病毒或病毒颗粒中。

24. 一种生产权利要求 17 的文库的方法，所述方法包括：

- (a) 提供一系列核酸，每个核酸均包含编码来自一天然存在的蛋白质的锌指结构域的序列；和
- (b) 将该系列的每个核酸与该系列的一或多个其它核酸连接以形成众多的嵌合核酸，其中所述嵌合核酸包括：i) 至少两个编码来自不同的天然存在的蛋白质的锌指结构域的序

列，或 ii) 至少两个编码来自同一的天然存在的蛋白质的锌指结构域且呈与天然构象所不同的构象的序列。

25. 权利要求 24 的方法，其中步骤 (a) 包括使用与编码保守的结构域边界的序列退火的寡核苷酸引物从基因组 DNA、一种信使 RNA (mRNA) 混合物、或者一种互补 DNA (cDNA) 混合物扩增编码锌指结构域的多核苷酸的集合。

26. 权利要求 24 的方法，其中步骤 (a) 包括：

- (i) 选择众多的锌指结构域，每个结构域均对感兴趣的靶位点内的一序列具有特异性；
- (ii) 提供众多的多核苷酸，每个多核苷酸均编码至少一个所选择的锌指结构域，从而提供所述多核苷酸系列。

27. 权利要求 26 的方法，其中所述众多的锌指结构域是通过查询包括将锌指结构域与其各自的结合位点相关联的信息的数据库而选择的。

28. 一种产生与一靶 DNA 位点特异性结合的人工锌指多肽的方法，所述方法包括：

- 提供权利要求 1 的多肽文库；
- 将靶 DNA 位点与该文库的多肽接触；和
- 鉴别与靶 DNA 位点特异性结合的一或多个多肽。

29. 权利要求 28 的方法，其中所述文库的多肽固定在一个固体支持物上。

30. 权利要求 28 的方法，其中所述文库的每个多肽均展示在病毒或病毒颗粒的表面。

31. 一种鉴别编码特异性识别一靶 DNA 位点的锌指多肽的核酸的方法，所述方法包括：

- 提供权利要求 17 的多核苷酸文库；
- 提供含有与靶 DNA 位点可操纵地连接的报告基因的细胞；
- 在细胞中表达众多的多核苷酸；
- 鉴别一种细胞，其中报告基因的表达水平相对于在没有识别靶

DNA 位点的多肽的情况下的表达水平发生变化；和

鉴别在该细胞中表达的一种多核苷酸，所述多核苷酸是众多多核苷酸之一，由此鉴别编码特异性识别靶 DNA 位点的多肽的多核苷酸。

32. 权利要求 31 的方法，其进一步包括修饰所鉴别的锌指多肽的氨基酸序列而不改变锌指多肽与靶 DNA 位点的结合特异性。

33. 权利要求 31 的方法，其中靶位点包含至少 6 个预定的核苷酸。

34. 权利要求 31 的方法，其中所述细胞是酵母细胞。

35. 权利要求 31 的方法，还包括将所述多核苷酸导入每一细胞的步骤。

36 权利要求 31 的方法，进一步包括将含有所述报道基因的细胞与包含所述文库的多核苷酸的细胞融合的步骤。

37. 一种多肽，其包含一个第一个和一个第二个锌指结构域，其中所述第一个和第二个锌指结构域独立地选自表 5、6 和 7 的锌指结构域；其中所述第一个和第二个锌指结构域各自对应于一种来自一天然存在的蛋白质的锌指结构域，且 (i) 不存在于该同一的天然存在的蛋白质中，或 (ii) 存在于该同一天然存在的蛋白质中但呈与其在所述多肽中的构象不同的构象。

38. 权利要求 37 的多肽，其进一步包含一个第三个锌指结构域，其中第一、第二和第三个锌指结构域系列示于表 10 的一行中。

39. 一种核酸序列，其包含编码权利要求 37 的多肽的多核苷酸。

40. 一种纯化的多肽，其包含选自以下一组的一个氨基酸序列：
SEQ ID NO:167, 168, 169, 170, 171, 172, 173, 175, 177, 179, 181, 183, 185, 187, 189, 191, 193, 195 和 196。

41. 一种核酸序列，其包含编码权利要求 40 的多肽的多核苷酸。

锌指结构域文库**发明领域**

本发明涉及 DNA 结合蛋白，如转录因子。

发明背景

大多数基因在转录水平由多肽转录因子调节，所述转录因子结合基因内典型地在启动子或增强子区域内的特异性 DNA 位点。这些蛋白质激活或阻抑 RNA 聚合酶在启动子处的转录起始作用，从而调节靶基因的表达。许多转录因子包括激活子和阻抑物的结构均是模块化的。这种模块可以折叠为结构独特的结构域并具有特异功能，如 DNA 结合，二聚体化或者与转录机制相互作用。效应结构域如激活结构域或阻抑结构域当转移至异源转录因子 DNA 结合区时保留其功能（Brent 和 Ptashne, (1985) 细胞 43:729-36; Dawson 等, (1995) 分子细胞生物学 15:6923-31）。从 NMR 和 X 线晶体学数据中确定了许多 DNA 结合区的三维结构，包括锌指结构域，同源域及螺旋-转角-螺旋结构域。

发明概述

本发明提供了一种快速的可扩展的基于细胞的方法，以鉴别和构建嵌合蛋白，例如转录因子。这种转录因子可用于例如在生物医学和生物工程学应用中改变内源基因的表达。所述转录因子在体内分析，即在完整的培养的活细胞中分析。从该分析结果中可以构建数据库，该数据库可用于设计其它嵌合的转录因子。可以制备嵌合转录因子的文库并筛选。所述嵌合蛋白还可以结合并调节除了 DNA 之外的分子例如蛋白质和 RNA、特别是小的非编码 RNA（ncRNA）。

一方面，本发明的特征是编码嵌合的锌指蛋白的核酸的文库。术语“文库”是指相似的但不相同的生物分子的物理集合。该集合可以

例如在一个容器中或者物理分离在一些独立的容器中（成组或单独地）或者位于一固体支持物上的独立的位置。该文库各个成员可以多份存在于该集合中。

第一个文库包括众多核酸 (a plurality of nucleic acids)，每个核酸均编码包含至少第一，第二和第三个锌指结构域的一种多肽。本文所用“第一，第二和第三个”是指三个独立的结构域，其可以以任何顺序在多肽中存在：例如每个结构域均可以出现在其它两个结构域中任一个或这两个结构域的 N 末端或 C 末端。第一个锌指结构域在所述众多核酸之中变化。第二个锌指结构域在所述众多核酸之中变化。在文库中存在至少 10 种不同的第一个锌指结构域。在一个实施方案中，该文库的至少 0.5%，1%，2%，5%，10%或 25%的成员具有以下之一或这两种性质：(1)每个成员均阻抑至少一个 p1G 报道质粒在体内的转录至少 1.25 倍；(2) 每个成员均结合至少一个靶位点，解离常数不超过 7, 5, 3, 2, 1, 0.5 或 0.05nM。第一个和第二个锌指结构域可以来自不同的天然存在的蛋白质，或者位于与其在天然存在的蛋白质中相关位置不同的构形中。例如，第一个和第二个锌指结构域在多肽中可以相邻，但在天然存在的蛋白质中可以由一或多个间隔锌指结构域分隔。

本文所用术语“解离常数”是指多肽与 28 个碱基对的双链 DNA 结合的平衡解离常数，所述 DNA 包括一个 9 碱基对的靶位点。解离常数通过凝胶移位分析使用一种纯化的蛋白质测定，所述蛋白质在室温在 20mM Tris pH7.7, 120mM NaCl, 5mM MgCl₂, 20μM ZnSO₄, 10%甘油, 0.1%Nonidet P-40, 5mM DTT 和 0.10mg/ml BSA（牛血清白蛋白）中是结合的。另外的详细描述见实施例 10 及 Rebar 和 Pabo（(1994) 科学 263:671—673）。

本文所用短语“阻抑 p1G 报道质粒的转录”是指 p1G 报道质粒的萤光素酶报道基因的阻抑倍数，所述报道质粒具有一个给定的 9 碱

基对的靶位点位于 TATA 区的下游，如图 12 所示。阻抑倍数通过实施例 68 所述方法测定，需要用合适的表达和报道质粒转染 HEK 细胞。

第一个和第二个锌指结构域可以是天然存在的结构域，例如下文所述。

第二个本发明的文库包括众多核酸，每个核酸均编码一种多肽，所述多肽包括至少第一个和第二个锌指结构域。每种多肽的第一个和第二个锌指结构域（1）与不同的天然存在的蛋白质的锌指结构域相同（并且通常不存在于相同的天然存在的蛋白质中或者呈与其在天然存在的蛋白质中相关位置不同的构形），（2）有不超过 4，3，2 或 1 个氨基酸残基与天然存在的蛋白质的结构域不同，或者（3）是来自天然存在的蛋白质的不相邻的锌指结构域。相同的锌指结构域是指从第一个金属配位残基（典型是半胱氨酸）至最后一个金属配位残基（典型是组氨酸）的每个氨基酸均相同的锌指结构域。第一个锌指结构域在所述众多核酸之中变化，第二个结构域在所述众多核酸之中变化。天然存在的蛋白质可以是任何真核锌指蛋白质：例如真菌（如酵母），植物或动物蛋白（例如哺乳动物蛋白，如人或鼠蛋白）。每种多肽均可进一步包括第三，第四，第五和/或第六个锌指结构域。每个锌指结构域均可以是哺乳动物例如人锌指结构域。典型地当第一个和第二个锌指结构域呈人工构形来自相同的天然存在的蛋白质时，先前不相邻的结构域此时相邻，先前的 N 末端是现在的 C 末端等。

针对上述第一个和第二个文库，每种多肽的第一个和/或第二个锌指结构域可以选自表 5，6 和 7。在另一个实施例中，至少一种多肽的第一个和/或第二个锌指结构域可以选自表 5，6 和 7。在一个实施方案中，至少 1%，5%，10%，25%，50%，75%或者所有的表 5，6 和 7 列出的锌指结构域是由所述众多核酸中的至少一个核酸编码的。由所述众多核酸的各成员编码的众多第一个锌指结构域，可以包括足够数目的不同的锌指结构域，以特异性结合至少 10，20，30，

40 或 50 个不同的 3 个碱基对的 DNA 位点。

如果包括 Zif268 的第 1 个和第 2 个锌指 (finger 1 or 2) 及一给定的锌指结构域的一种嵌合蛋白与下述靶位点的亲和性为至少 5nM, 则认为该给定的锌指结构域“特异性结合”一个给定的 3 个碱基对的 DNA 位点, 所述靶位点包括该给定的 3 个碱基对的 DNA 位点及由 Zif268 的第 1 个和第 2 个锌指识别的 5bp 序列 5'-GGGCG-3'。Zif268 的第 1 个和第 2 个锌指具有以下多肽序列: ERPYACPVE SCDRRFSR SDELTRHIRIHTGQKPFQCRICMRNFSRSDHLTTHIRTH (SEQ ID NO: 198)。术语“识别”和“特异性结合”可交替使用, 是指在上述 Zif268 融合分析中对锌指结构域的结合位点的辨别。

任选地, 所述众多核酸集合编码足够数目的不同的锌指结构域, 以识别至少 10, 20, 30, 40 或 50 个不同的 3bp DNA 位点。在一个实施方案中, 所述众多核酸集合编码足够数目的不同的锌指结构域, 以识别不超过 40, 30, 20, 10 或 5 个不同的 3bp DNA 位点。

所述众多核酸可以集合编码至少 5, 10, 20, 30 或 40 个不同的第一个锌指结构域和/或至少 5, 10, 20, 30 或 40 个不同的第二个锌指结构域。所述众多核酸可包括至少 10, 50, 200, 500, 1000, 5000, 10000, 20000, 25000 或 40000 个不同的核酸 (即具有不同的序列)。在一些情况中, 众多的核酸可包括不超过 100, 500, 2000, 5000, 15000, 30000 或 50000 个核酸。众多的核酸以摩尔比率可组成至少 20%, 50%, 70%, 80%, 90%, 95% 或 100% 的文库。

在一个实施方案中, 由所述众多核酸编码的多肽包括不同数目的锌指结构域。例如, 由第一个亚群 (subset) 编码的多肽可以包括 4 个锌指结构域, 由第二个亚群编码的多肽可以包括 5 个锌指结构域。另一种组合是具有 3、4 和 5 个结构域, 或者 4、5 和 6 个结构域。

在一个实施方案中, 由所述众多核酸编码的多肽包括不同类型的转录调节结构域。例如, 由第一个亚群编码的多肽可包括一个转录激

活结构域，由第二个亚群编码的多肽可包括一个转录阻抑结构域。另一个亚群编码的多肽可无转录调节结构域。这个实施方案使得能不基于特定类型的转录因子而筛选文库。

在一个实施方案中，每个核酸均固定于一固体支持物上。在另一个实施方案中，每个核酸均附着于其编码的多肽。这种附着可以是共价或非共价的。例如，由每个核酸编码的多肽可以附着于病毒或病毒颗粒的外部，及所述核酸包装在病毒或病毒颗粒内部。“病毒”是指可以感染宿主细胞并复制的遗传实体。“病毒颗粒”是指可以感染宿主细胞但不能复制的遗传实体。病毒颗粒例如是丝状噬菌体外包被，其包括一种噬菌粒核酸。病毒或病毒颗粒可以感染哺乳动物细胞（如反录病毒或腺病毒）或者细菌细胞（例如噬菌体）。在另一个实施例中，所述多肽例如通过嘌呤霉素连接共价附着于核酸。

在另一个实施例中，每个多肽还包括一个激活或阻抑结构域。每个多肽的第一个和第二个锌指结构域可以彼此相邻或者由一个间隔结构域或接头分离。

在一个实施方案中，所述文库的每个核酸均在细胞中。所述核酸可以在细胞内表达。所述细胞还可以包括一个异源报道基因构建体，其包括可操纵地与报道基因连接的一个靶 DNA 位点。所述细胞可以是酵母细胞，动物细胞如哺乳动物细胞，鸟类细胞或昆虫细胞；细菌细胞或植物细胞。

所述文库的所有核酸均可以位于一个单一容器内或者在一个单一表面上。在另一个实施方案中，所述众多核酸的亚群位于独立的容器内，位于独立的表面上，或者位于同一表面的各独立部分上。在再一个实施方案中，所述文库的每个核酸是可寻址的（addressable），例如可唯一寻址。“可寻址”的元件位于限定的空间位置，所述位置可以在适当条件下访问以检索（retrieve）该元件。例如，每个核酸均可位于微滴定平板的孔中，位于平面阵列上，或者位于冷冻的细胞样

品中。

所述文库的每个核酸均可以通过与关于所述核酸编码的多肽识别靶位点的能力的信息相关的指示器（pointer）在机读介质中查询。所述信息可以包括例如表示多肽与靶位点的结合亲和性的数值，预测多肽识别靶位点的能力的数值，或者反映多肽对内源基因在细胞（例如人体细胞）内表达的作用的一系列数值。

相似地，本发明提供了多肽的文库。所述文库包括众多多肽，其均由所述核酸文库的核酸编码。所述多肽文库也可包括在此所述核酸文库的任何合适特征，除了所述特征在蛋白质水平具体表现之外。

本发明还提供了一个试剂盒，其包括所述文库和一种机读介质，所述机读介质包括编码在其中的多肽识别靶 DNA 位点的能力的信息，其中所述文库的每个核酸或多肽均可以通过与关于核酸编码的各自多肽识别靶 DNA 位点的能力的信息相关的指示器在机读介质中查询。例如，所述信息可以包括表示多肽与靶位点结合亲和性的数值，或者预测多肽识别靶位点的能力的数值。所述试剂盒还包括电脑可读说明书，其使用户可接触所述信息。

在另一方面，本发明涉及一种多肽，其包括第一个和第二个锌指结构域。每个锌指结构域具有选自表 5，6 和 7 的序列。第一个和第二个锌指结构域来自不同的天然存在的多肽。所述多肽还可包括第三，第四和第五个锌指结构域。这些结构域的每一个还可以具有选自表 5，6 和 7 的序列。典型地，所述锌指结构域彼此相邻以形成锌指结构域阵列。这种阵列是一个多肽单位，其不由其它类型的结构或功能蛋白质结构域间断。

本发明还涉及一种多肽，其包括第一、第二和第三个锌指结构域。每个结构域均是天然存在的。至少两个结构域存在于不同的天然存在的多肽中。另外，所述多肽具有以下一或这两种性质：（1）在体内阻抑至少一个 p1G 报道质粒的转录至少 1.25，1.5，1.7，1.9，2.0 或 2.5

倍；(2) 结合至少一个靶核酸位点，解离常数不超过 7, 5, 3, 2, 1, 0.5 或 0.05nM。所述靶位点可以是 DNA 或 RNA。所述第一、第二和/或第三个锌指可选自表 5, 6 和 7。在一个实施方案中，每个给定多肽的第一、第二和第三个锌指结构域由表 10 的一行中列出的结构域表示，例如第 113 行之上的一行，例如包括所有人体锌指结构域的一行。在一个实施方案中，所述多肽仅包括来自天然存在的多肽的结构域（例如哺乳动物例如人多肽）。

在一个实施方案中，第一个结构域位于第二个结构域的 N 末端，第二个结构域位于第三个结构域的 N 末端，第二个结构域发生在与第一个结构域不同的天然存在的多肽中。第一、第二和第三个结构域中的每一个可以存在于与其它两个不同的天然存在的多肽中。

关于任何所述多肽，其可进一步包括一个异源序列，例如核定位信号，小分子结合结构域（例如类固醇结合结构域），附加表位或纯化柄（handle），催化结构域（例如核酸修饰结构域，核酸切割结构域，或者 DNA 修复催化结构域），转录功能结构域（例如激活结构域，阻抑结构域等），蛋白质转导结构域（例如来自 HIV tat），和/或调节位点（例如磷酸化位点，蛋白化位点，或者蛋白酶切割位点）。

所述多肽可（共价或非共价）附着于固体支持物，例如珠，基质或平面阵列。所述多肽也可附着于一个标记如放射性化合物，荧光化合物，另一个可检测实体，或者检测系统的一种成分（例如化学发光剂）。

本发明还包括编码前述多肽的一种分离的核酸序列。所述核酸进一步包括一种可操纵地连接的调节序列，例如启动子，转录增强子，5'非翻译区，3'非翻译区，病毒包装序列和/或可检测标记。所述核酸可包装入病毒中，例如可感染哺乳动物细胞的病毒，例如慢病毒，反录病毒，痘病毒或腺病毒。

本发明还提供了含有所述核酸的细胞。所述细胞可以在个体生物

体的或培养的组织内。所述细胞可以是动物（例如哺乳动物），植物，或微生物（例如真菌或细菌）细胞。本发明还包括一种非人转基因哺乳动物，例如小鼠，大鼠，猪，兔，牛，山羊或绵羊。转基因哺乳动物的遗传互补体包括编码上述或本文其它部分所述的嵌合锌指多肽的核酸序列。本发明还包括生产所述多肽的方法，例如通过表达核酸进行，及使用所述多肽的方法，例如调节细胞中的内源基因或病毒基因。

再一方面，本发明涉及评价由核酸文库编码的一或多种多肽的方法。所述方法包括：提供文库，其中众多文库核酸中的每一个核酸均位于细胞内（在任何给定的细胞中具有一或多个（但非全部）文库成员）；在其存在的细胞中表达所述众多核酸的每一个核酸；鉴别相对于无文库成员的细胞而具有改变的报道基因表达的细胞，从而鉴别编码识别靶 DNA 位点的多肽的核酸。

另一方面，本发明涉及构建嵌合锌指结构域文库的方法。所述方法包括：提供一系列核酸，每种核酸均包含编码选自表 5，6 和 7 所示锌指结构域的序列；将所述系列的每一种核酸与该系列的一或多种，优选 2，3 或 4 种其它核酸结合以形成众多嵌合核酸。每种嵌合核酸均可位于载体中，例如哺乳动物表达载体中。

在一个实施方案中，所述方法进一步包括在结合之后，将众多的嵌合核酸的一或多个导入细胞中并表达所述一或多个核酸。在另一个实施方案中，所述方法还包括将一或多个嵌合核酸逐个导入细胞例如哺乳动物细胞中；表达所述嵌合核酸；监测基因或蛋白质在细胞中的表达。所述细胞可包括一个报道基因构建体，例如包括据信由嵌合核酸编码的多肽识别的靶位点的构建体，由此所述多肽结合靶位点并诱导或阻抑报道基因的表达。

再一方面，本发明涉及鉴定一种嵌合锌指蛋白例如本文所述锌指蛋白的方法。所述方法包括：将编码所述蛋白质的核酸导入细胞中，

表达所述核酸，确定内源基因在细胞中的表达分布图（profile）。这种表达分布图包括众多数值，其中每个数值相当于不同的基因、基因的剪接变体或等位基因变体的表达水平（即 mRNA 水平）或者翻译产物的丰度（即蛋白质水平）。该数值可以是对基因或者基因的翻译产物的表达水平进行的定量或定性测定，即测定 1)从基因中转录的 mRNA 或 2)由该基因编码的多肽的丰度。

所述方法还可包括将测定的表达分布图与至少一个参照（reference）表达分布图相对比，从而鉴定嵌合的锌指蛋白。所述参照分布图可以是没有异源嵌合锌指蛋白或者包括对照载体的一个相关细胞的表达分布图。所述对比可以鉴别由嵌合锌指蛋白而改变的一或多个基因的调节。在一个实施方案中，将样品表达分布图与参照分布图对比以产生差异分布图。样品表达分布图也可以在多维空间与一组参照分布图对比。在一个实施方案中，样品表达分布图使用一个核酸阵列确定。在另一个实施方案中，样品表达分布图使用不需要阵列的方法和/或设备（例如 SAGE 或使用多个引物的定量 PCR）测定。

所述方法还可包括确定或指定多肽的靶结合位点，并鉴别其调节已由多肽改变的基因的调节核酸序列中靶结合位点的出现情况。所述方法还可用于区分直接和间接靶。

另一方面，本发明涉及一种方法，包括：提供众多的核酸，每个核酸均编码包含第一个和第二个锌指结构域的多肽，其中由每个核酸编码的多肽的第一个和第二个锌指结构域与来自不同的天然存在的哺乳动物蛋白的锌指结构域相同，第一个锌指结构域在所述众多核酸之中变化，第二个锌指结构域在所述众多核酸之中变化；将每个核酸均导入在导入之前具有一给定特性的细胞中，从而提供众多细胞；在所述众多细胞的每个细胞中表达导入的核酸；从所述众多细胞中鉴别其中给定特性改变的细胞。所述方法可包括本文所述其它特征。所述特性例如包括对一种情况（例如应激）的敏感性或抗性增强，增殖能

力改变,致病性改变,产物产生(例如代谢物产生)改变。

再一方面,本发明涉及一种鉴别可结合特定靶位点的嵌合锌指蛋白的方法。所述方法包括:提供数据记录,每个记录涉及一个天然存在的人锌指结构域的标识符(identifier)及至少一个3或4bp的亚位点,所述亚位点由标识符指代的锌指结构域识别;将靶位点分解(parsing)成至少两个3或4bp的亚位点;针对每个亚位点,从数据记录中检索出一系列标识符,包括识别亚位点的锌指结构域的标识符;设计一种多肽,其包含针对每个亚位点的一个锌指结构域,该锌指结构域由各自亚位点的标识符指代。

所述数据记录可包括鉴别选自表5,6和7所示人锌指结构域的一种记录。所述方法可进一步包括合成编码所述多肽的核酸和/或在体外合成所述多肽的步骤。所述方法还可包括测定所述多肽与靶位点结合的步骤,例如使用体外结合分析或体内分析如报道基因表达分析确定。所述合成的多肽可进一步包括一个激活或阻抑结构域。

在一个实施方案中,所述方法进一步包括测定多肽改变一或多个内源基因表达的能力。该测定可包括例如使用核酸微阵列作出多个内源基因表达的分布图。所述方法也可进一步包括例如在体外将所述多肽与包括靶位点的DNA接触。

在另一个实施方案中,所述方法进一步包括从可寻址的核酸文库中检索编码所述多肽的核酸,文库的每个核酸均包含编码第一个和第二个锌指结构域的序列。

本发明还涉及一种方法,包括:将数据记录存储于机读介质中,每个记录均涉及一个锌指结构域标识符及由标识符所指代的锌指结构域识别的至少一个3或4bp的亚位点;从存储记录中检索出与感兴趣的亚位点相关的一或多个标识符;构建编码多肽的核酸,该多肽包含(a)由一或多个检索出的标识符之一指代的锌指结构域,(b)第二个DNA结合结构域。第二个DNA结合结构域可以是锌指结构域。

所述构建可包括构建众多核酸，每个核酸均包含编码检索的标识符指代的锌指结构域的序列。所述方法可进一步包括针对每个核酸，在细胞中表达所述核酸，测定与没有所述核酸或所述核酸未表达时给定基因的转录水平相比，当所述核酸被表达时给定基因转录水平的变化。所述测定可进一步包括通过作出分布图测定多个基因的转录水平。

另一方面，本发明涉及一种基于计算机的方法，包括：存储包括（a）与（b）之间的关联的信息，其中（a）为表 5，6 和 7 所示天然存在的众多锌指结构域的每一个结构域，（b）为由所述结构域识别的一或多个亚位点；接受用户查询，查询包括指明靶核酸序列的字符串；过滤信息以鉴别预测识别靶核酸序列内一个位点的锌指结构域组合。

所述方法可进一步包括将所述组合展示给用户或者物理定位位于一个文库核酸或多肽，所述文库核酸或多肽包括从可寻址的核酸或多肽文库鉴别的锌指结构域组合之一。

再一方面，本发明涉及存储于机读介质上的一个数据库。该数据库包括：（i）数据，其代表（a）各个天然存在的锌指结构域，（b）核酸位点，（c）包含众多天然存在的锌指结构域的嵌合多肽；（ii）关于（1）各个锌指结构域与各个结构域识别的核酸位点，（2）嵌合多肽与其各自锌指结构域成分，（3）嵌合多肽与由各自嵌合多肽识别的核酸位点，之间的关系。所述数据库使用户可以鉴别预计识别靶核酸序列内位点的锌指结构域组合。

所述数据还包括（d）可寻址位置，该位置与阵列的多肽文库中驻在的嵌合多肽相关，或者（e）如上所述的表达分布图。每个表达分布图可以与一嵌合多肽相关联。

本发明还涉及一个文库，其包括众多的多肽，每个多肽均包含第一个和第二个锌指结构域，其中每个多肽的第一个和第二个锌指结构域与来自不同的天然存在的多肽的哺乳动物锌指结构域相同，第一个

锌指结构域在众多多肽之中变化，第二个锌指结构域在众多多肽之中变化。文库的每个多肽均可附着于一固体支持物（例如珠，基质或平面阵列）。

本发明还提供了一种对测试核酸作出分布图（profiling）的方法。该方法包括：将测试核酸与所述文库的多肽接触；鉴别特异性结合测试核酸的一或多个多肽。所述多肽可固定于可寻址阵列上或附着于病毒颗粒。

本发明还涉及一种鉴别识别 DNA 上靶位点的肽结构域的方法。这个方法在本文有时称为“结构域选择方法”或“体内筛选方法”。该方法包括提供：（1）含有报道基因构建体的细胞，（2）众多的杂合核酸。所述报道基因构建体具有可操纵地与启动子连接的报道基因，所述启动子既具有募集位点又具有靶位点。当转录因子识别（即超过背景水平的结合）启动子的募集位点和靶位点这两个位点时，报道基因的表达高于给定水平，但当转录因子仅识别启动子的募集位点时，报道基因的水平不高于给定水平。每个杂合核酸均编码具有以下元件的非天然存在的蛋白质：（i）转录激活结构域，（ii）识别募集位点的 DNA 结合结构域，（iii）测试锌指结构域。测试锌指结构域的氨基酸序列在众多的杂合核酸的成员之间变化。所述方法还包括：在使至少一个所述核酸进入至少一个细胞的条件下将所述众多核酸与所述细胞接触；在使杂合核酸在细胞中表达的条件下维持细胞；鉴别以高出给定水平表达报道基因的细胞，作为细胞含有编码识别靶位点的测试锌指结构域的杂合核酸的指征。

DNA 结合结构域即识别募集位点及在众多成员之间不变化的结构域，可包括例如 1，2，3 或多个锌指结构域。所述方法中利用的细胞可以是原核或真核细胞。真核细胞例如是酵母细胞，如 *Saccharomyces cerevisiae*，*Schizosaccharomyces pombe*，或者 *Pichia pasteuris*；昆虫细胞如 Sf9 细胞；哺乳动物细胞如成纤维细胞或淋巴

细胞。

“给定水平”是当转录因子识别募集位点但不识别靶位点时观测的表达量。在一些情况中，“给定水平”可以是 0（至少在所用分析的检测限度内）。

所述方法可包括一个额外的步骤，即从核酸如基因组 DNA，mRNA 混合物或 cDNA 混合物中扩增编码测试锌指结构域的源核酸以产生扩增的片段。所述源核酸可以使用寡核苷酸引物扩增。所述寡核苷酸引物可以是一系列简并寡核苷酸之一（例如具有不同核酸序列的特异寡核苷酸集合，或者具有非天然碱基如肌苷的特异的寡核苷酸），其与编码保守的结构域边界的核酸退火。或者，所述引物可以是特异的寡核苷酸。利用扩增的片段产生杂合核酸以包含于前述方法中使用的众多杂合核酸中。

所述方法可进一步包括以下步骤：(i) 在序列数据库中鉴别候选的锌指结构域氨基酸序列；(ii) 提供编码候选锌指结构域氨基酸序列的一个候选核酸；(iii) 利用该候选核酸构建杂合核酸以包含于前述方法中使用的众多杂合核酸中。所述数据库可包括多个氨基酸序列的记录例如已知和/或推测的蛋白质，以及多个核酸序列的记录如 cDNA，EST，基因组 DNA 或经计算处理除去预测的内含子的基因组 DNA。

如果需要，可以重复所述方法以鉴别识别第二个靶位点的第二个测试锌指结构域，所述第二个靶位点例如是除了第一个测试锌指结构域识别的位点之外的位点。随后，可以构建编码第一个和第二个鉴别的测试锌指结构域的核酸。编码的杂交蛋白特异性识别靶位点，所述靶位点包括第一个测试锌指结构域的靶位点和第二个测试锌指结构域的靶位点。

本发明还涉及一种确定测试锌指结构域是否识别启动子上靶位点的方法。这个方法在本文有时称为“位点选择方法”。所述方法包

括提供报道基因构建体和杂合核酸的步骤。所述报道基因可操纵地与包括一个募集位点和一个靶位点的启动子连接，并且当转录因子既识别启动子的募集位点又识别靶位点时以超出给定水平表达，但当转录因子仅识别启动子的募集位点时则不以超出给定水平表达。所述杂合核酸编码具有以下元件的非天然存在的蛋白质：(i) 一个转录激活结构域，(ii) 一个识别募集位点的 DNA 结合结构域，(iii) 一个测试锌指结构域。所述方法还包括：在使报道基因构建体进入细胞的条件下将报道基因构建体与细胞接触；在前述步骤之前，之后或同时在使杂合核酸进入细胞的条件下将杂合核酸与细胞接触；在使杂合核酸在细胞中表达的条件下维持细胞；检测细胞中报道基因表达。报道基因表达水平高于给定水平是测试锌指结构域识别靶位点的指征。

所述报道基因构建体和杂合核酸可以包含于独立的质粒中。这两个质粒可以同时或连续导入细胞中。这两个质粒之一或两者可含有选择标记。所述报道基因构建体和杂合核酸也可以包含于同一质粒上，在这种情况下只需要一个接触步骤将两个核酸导入细胞中。在另一个实施方案中，将这两个核酸之一或两者稳定整合入细胞的基因组中。对于这种方法，与在此所述任何体内方法一样，所述转录激活结构域可以用转录阻抑结构域置换，并鉴别报道基因的表达水平降低至给定水平之下的细胞。

本发明的另一种方法是通过融合两种细胞而便于快速确定测试锌指结构域的结合性质。所述方法包括：提供含有报道基因的第一种细胞；提供含有杂合核酸的第二种细胞；将第一种和第二种细胞融合以形成融合细胞；在使杂合核酸在融合的细胞中表达的条件下维持融合的细胞；检测融合的细胞中报道基因的表达，其中报道基因的表达水平高于给定水平表明测试锌指结构域识别靶位点。例如，第一和第二种细胞可以是组织培养细胞或真菌细胞。所述方法的一个实施方案是利用 *S.cerevisiae* 细胞。第一种细胞具有第一种交配型，例如 MATa；

第二种细胞具有不同于第一种交配型的第二种交配型，例如 MAT α 。将这两种细胞彼此相接触，酵母接合产生一个单细胞(例如 MATa/ α)，其具有同时含有第一和第二种细胞的基因组的细胞核。所述方法可包括提供具有相同第一种交配型的多个第一种细胞，其中每个第一种细胞具有一个报道基因构建体，所述构建体具有不同的靶位点；提供具有相同第二种交配型的多个第二种细胞，每个细胞具有不同的测试锌指结构域。产生一个多种成对方式交配例如所有可能的成对交配的矩阵。所述方法用于确定多个测试锌指结构域对多个结合位点的结合优先性，所述结合位点例如是一套完整的可能靶位点。

本发明还提供了一种分析测试锌指结构域的结合优先性的方法。所述方法包括提供(1)细胞，基本上所有细胞均含有杂合核酸，(2)众多的报道基因构建体。每个报道基因构建体均具有可操纵地与具有募集位点和靶位点的启动子连接的报道基因。当转录因子既识别启动子的募集位点又识别靶位点时，报道基因以超出给定水平表达，但当转录因子仅结合启动子的募集位点时，报道基因不以超出给定水平表达。第二个靶位点在众多报道基因构建体的成员之间变化。所述杂合核酸编码具有以下元件的一种杂合蛋白：(i) 转录激活结构域，(ii) 识别募集位点的 DNA 结合结构域，(iii) 测试锌指结构域。所述方法还包括：在使至少一种报道基因构建体进入至少一个细胞中的条件下将众多的报道基因构建体与细胞接触；在使所述核酸在所述细胞中表达的条件下维持细胞；鉴别一种细胞，其含有报道基因构建体并以超出给定水平表达所述报道基因构建体，这是细胞中报道基因构建体具有所述锌指结构域识别的靶位点的指征。

如果测试锌指结构域对一个以上靶位点具有结合优先性(binding preference)，可以通过上述方法鉴别每个细胞均具有不同靶位点的众多细胞。所述方法可进一步包括鉴别呈现最高水平报道基因表达的细胞。或者，测定报道基因的极限表达水平，例如报道基因表达提高2，

4, 8, 20, 50, 100, 1000 倍或更高, 并选择在报道基因在极限水平之上表达的所有细胞。

靶结合位点的长度例如可以是 2-6 个核苷酸。众多的报道基因构建体可以包括 A, T, G 和 C 核苷酸在靶结合位点的 2, 3 或 4 或多个位置的每一种可能组合。

另一方面, 本发明涉及一种鉴别众多锌指结构域的方法。所述方法包括: 进行结构域选择方法以鉴别第一种测试锌指结构域, 再次进行结构域选择方法以鉴别第二种测试锌指结构域, 第二种测试锌指结构域识别与第一种测试结构域的靶位点不同的靶位点。本发明还涉及一种产生编码嵌合锌指蛋白的核酸的方法, 所述方法包括进行两次结构域选择方法以鉴别第一种和第二种测试锌指结构域, 及构建编码一种多肽的核酸, 所述多肽包括第一种和第二种测试锌指结构域。所述核酸可编码一种杂交蛋白, 其包括特异性识别包括两个亚位点的位点的两个结构域。所述亚位点是第一种测试锌指结构域的靶位点和第二种测试锌指结构域的靶位点。可重复进行所述方法以鉴别另外的锌指结构域, 并构建编码一种多肽的核酸, 所述多肽包括 3, 4, 5, 6 或多个锌指结构域以特异性识别核酸结合位点。

再一方面, 本发明涉及一种鉴别由锌指结构域识别的 DNA 序列的方法。所述方法包括: 进行位点选择方法以鉴别对第一种测试锌指结构域的第一种结合优先性, 并再次进行位点选择方法以鉴别对第二种测试锌指结构域的第二种结合优先性。可以构建一种核酸, 其编码第一和第二种鉴别的测试锌指结构域。所述核酸可编码一种杂合蛋白, 其包括特异性识别一个位点的两种结构域, 所述位点包括第一种测试锌指结构域的靶位点和第二种测试锌指结构域的靶位点。可重复进行所述方法以鉴别另外的锌指结构域, 并构建编码一种多肽的核酸, 所述多肽包括 3, 4, 5, 6 或多个锌指结构域以特异性识别核酸结合位点。

本发明还涉及一种鉴别识别 DNA 上靶位点的肽结构域的方法。所述方法包括提供 (1) 含有报道基因构建体的细胞, (2) 众多的杂合核酸。所述报道基因构建体具有可操纵地与启动子连接的报道基因, 所述启动子既具有募集位点又具有靶位点。当转录因子识别 (即以超出背景水平结合) 启动子的募集位点和靶位点这两个位点时, 所述报道基因以低于给定水平表达, 但当转录因子仅识别启动子的募集位点时, 报道基因不以低于给定水平表达。每个杂合核酸均编码具有以下元件的非天然存在的蛋白质: (i) 转录阻抑结构域, (ii) 识别募集位点的 DNA 结合结构域, (iii) 测试锌指结构域。测试锌指结构域的氨基酸序列在众多杂合核酸的成员之间变化。所述方法进一步包括: 在使至少一种核酸进入至少一个细胞的条件下将众多的核酸与细胞接触; 在使杂合核酸在细胞中表达的条件下维持细胞; 鉴别以低于给定水平表达报道基因的细胞, 这是细胞含有编码识别靶位点的测试锌指结构域的杂合核酸的指征。这个方法的另外实施方案与利用转录激活结构域的方法相似。另外, 使用转录阻抑结构域代替转录激活结构域可以进行在此所述的任何其它选择方法。

另一方面, 本发明涉及某些纯化的多肽和分离的核酸。本发明的一种纯化多肽可包括具有以下一或多个氨基酸序列的多肽:

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Cys-X-Ser-Asn-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:68),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-His-X-Ser-Asn-X_b-X-Lys-His-X_{3.5}-His (SEQ ID NO:69),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Ser-X-Ser-Asn-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:70),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ser-Thr-X_b-X-Val-His-X_{3.5}-His (SEQ ID NO:71),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Val-X-Ser-X_c-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:72),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ser-His-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:73),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ser-Asn-X_b-X-Val-His-X_{3.5}-His (SEQ ID NO:74),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ser-X_c-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:75),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ala-His-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:150),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Phe-Asn-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:151),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ser-His-X_b-X-Thr-His-X_{3.5}-His (SEQ ID NO:152),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ser-His-X_b-X-Val-His-X_{3.5}-His (SEQ ID NO:153),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ser-Asn-X_b-X-Ile-His-X_{3.5}-His (SEQ ID NO:154),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ser-Asn-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:155),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Thr-His-X_b-X-Gln-His-X_{3.5}-His (SEQ ID NO:156),

Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Thr-His-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:157),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Arg-X-Asp-Lys-X_b-X-Ile-His-X_{3.5}-His (SEQ ID NO:158),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Arg-X-Ser-Asn-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:159),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Gly-Asn-X_b-X-Arg-His-X_{3.5}-His (SEQ ID NO:161),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Arg-X-Asp-Glu-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:162),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Arg-X-Asp-His-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:163),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Arg-X-Asp-His-X_b-X-Thr-His-X_{3.5}-His
(SEQ ID NO:164),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Arg-X-Asp-Lys-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:165),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Arg-X-Ser-His-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:166),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Arg-X-Thr-Asn-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:160),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-His-X-Ser-Ser-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:167),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Ile-X-Ser-Asn-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:168),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Lys-X-Ser-Asn-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:169),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ser-Asn-X_b-X-Lys-His-X_{3.5}-His
(SEQ ID NO:170),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Gln-X-Ser-His-X_b-X-Thr-His-X_{3.5}-His
(SEQ ID NO:171),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Val-X-Ser-Asn-X_b-X-Val-His-X_{3.5}-His
(SEQ ID NO:172),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Asp-X-Ser-Cys-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:193),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Ile-X-Ser-Asn-X_b-X-Val-His-X_{3.5}-His
(SEQ ID NO:194),

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Trp-X-Ser-Asn-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:195), or

X_a-X-Cys-X_{2.5}-Cys-X₃-X_a-X-Asp-X-Ser-Ala-X_b-X-Arg-His-X_{3.5}-His
(SEQ ID NO:196),

其中 X_a 是苯丙氨酸或酪氨酸, X_b 是疏水性残基, X_c 是丝氨酸或苏氨酸。本发明的核酸包括编码前述多肽的核酸。在一个实施方案中, 上述氨基酸序列是天然存在的序列。

另外, 本发明纯化的多肽可具有与以下序列至少 50%, 60%, 70%, 80%, 90%, 93%, 95%, 96%, 98%, 99% 或 100% 相同的氨基酸序列: SEQ ID NO:23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 103, 105, 107, 111, 113, 115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135, 137, 141, 143, 145, 147, 149, 173, 175, 177, 179, 181, 183, 185, 187, 189 或 191。所述多肽在相应于所述多肽的核酸接触残基的氨基酸位置可以与以下序列相同: SEQ ID NO:23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 103, 105, 107, 111, 113, 115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135, 137, 141, 143, 145, 147, 149, 173, 175, 177, 179, 181, 183, 185, 187, 189 或 191。或者所述多肽在相应于所述多肽的核酸接触残基的至少一个残基与以下序列不同: SEQ ID NO:23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 103, 105, 107, 111, 113, 115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135, 137, 141, 143, 145, 147, 149, 173, 175, 177, 179, 181, 183, 185, 187, 189 或 191。所述多肽也可以在除了 DNA 接触残基之外的至少一个残基处不同 (见以下进一步阐述)。例如在给定的锌指结构域内, 所述多肽可以有一个单氨基酸与上述氨基酸序列不同, 或者有 2, 3 或 4 个氨基酸与上述氨基酸序列不同。所述不同可由于在此所述的保守取代所致。在一个实施方案中, 与上述序列不同的氨基酸位于第二个锌配位半胱氨酸与 -1 DNA 接触位置 (参见下述 DNA 接触位置的编号系统) 之间。序列对比及两个序列之间

相同性百分率的测定可以使用数学公式完成。特别地，两个氨基酸序列之间相同性百分率使用 Needleman 和 Wunsch 算法测定 ((1970), 分子生物学杂志 48:444-453), 所述算法已经掺入 GCG 软件包的 GAP 程序中, 使用 Blossum 62 计分矩阵, gap penalty 为 12, gap extend penalty 为 4, 移码 gap penalty 为 5。

所述纯化的多肽也可以包括以下一或多种结构: 异源 DNA 结合结构域, 核定位信号, 小分子结合结构域 (例如类固醇结合结构域), 附加表位或纯化柄, 催化结构域 (例如核酸修饰结构域, 核酸切割结构域, 或者 DNA 修复催化结构域) 和/或转录功能结构域 (例如激活结构域, 阻抑结构域等)。在一个实施方案中, 所述多肽进一步包括第二个锌指结构域, 例如具有所述序列的结构域。例如, 所述多肽可包括锌指的一个阵列, 其包括两或多个锌指结构域。每个结构域均可具有选自以下一组的序列: SEQ ID NO:68-75, 150-172, 193-196 或者其亚群。另外, 每个结构域均可具有选自以下一组的一个序列: SEQ ID NO:23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 103, 105, 107, 111, 113, 115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135, 137, 141, 143, 145, 147, 149, 173, 175, 177, 179, 181, 183, 185, 187, 189 和 191 及其亚群。

如本文所述, 所述多肽可以在细胞中生产, 而且可以在细胞中通过结合靶位点而调节基因例如内源基因, 所述靶位点例如是包括各锌指结构域识别的亚位点的位点。见例如表 5, 6 和 7。

本发明还包括编码前述多肽的分离的核酸序列, 及在高度严格杂交条件下与一单链探针杂交的分离的核酸序列, 所述探针的序列由以下序列组成: SEQ ID NO:22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 102, 104, 106, 110, 112, 114, 116, 118, 120, 122, 124, 126, 128,

130, 132, 134, 136, 140, 142, 144, 146, 148, 150, 174, 176, 178, 180, 182, 184, 186, 188, 190, 192 或其互补序列。本发明还包括一种在细胞中表达与异源核酸结合结构域融合的本发明多肽的方法。所述方法包括将编码前述融合蛋白的核酸导入细胞中。本发明的核酸可通过异源核酸序列可操纵地调节, 所述异源核酸序列例如是诱导型启动子(例如类固醇激素调节的启动子, 小分子调节的启动子, 或者工程化的可诱导系统如四环素 Tet-On 和 Tet-Off 系统)。

术语“碱基接触位置”, “DNA 接触位置”或者“核酸接触位置”是指锌指结构域的 4 个氨基酸的位置, 其在结构上相应于 SEQ ID NO:21 的第 73 位精氨酸, 第 75 位天冬氨酸, 第 76 位谷氨酸和第 79 位精氨酸。这些位置也分别称为位置—1, 2, 3 和 6。为在查询序列中鉴别相应于碱基接触位置的位置, 将查询序列与感兴趣的锌指结构域排列对比, 由此查询序列的半胱氨酸和组氨酸残基与 Zif268 的第 3 指的那些残基排列对比。欧洲生物学信息协会的 ClustalW WWW Service (Thompson 等, (1994) 核酸研究 22:4673—4680) 提供了一种序列对比的常规方法。

保守氨基酸取代是指具有相似侧链的残基的互换性。例如, 具有脂族侧链的一组氨基酸是甘氨酸, 丙氨酸, 缬氨酸, 亮氨酸和异亮氨酸; 具有脂族羟基侧链的一组氨基酸是丝氨酸和苏氨酸; 具有含酰胺侧链的一组氨基酸是天冬酰胺和谷氨酰胺; 具有芳香族侧链的一组氨基酸是苯丙氨酸, 酪氨酸和色氨酸; 具有碱性侧链的一组氨基酸是赖氨酸, 精氨酸和组氨酸; 具有酸性侧链的一组氨基酸是天冬氨酸和谷氨酸; 具有含硫侧链的一组氨基酸是半胱氨酸和甲硫氨酸。根据情况, 同一组内的氨基酸可以互换。一些另外的保守氨基酸取代组是: 缬氨酸—亮氨酸—异亮氨酸; 苯丙氨酸—酪氨酸; 赖氨酸—精氨酸; 丙氨酸—缬氨酸; 天冬氨酸—谷氨酸; 及天冬酰胺—谷氨酰胺。

术语“异源多肽”是指具有非天然存在的序列的多肽(例如杂合

多肽), 或者具有与天然存在的多肽相同的序列但存在于非天然存在的周围环境中的多肽。

术语“杂合体”是指非天然存在的多肽, 其包含得自以下的氨基酸序列: (i) 至少两个不同的天然存在的序列; (ii) 至少一个人工序列 (即非天然存在的序列) 和至少一个天然存在的序列, 或者 (iii) 至少两个人工序列 (相同或不同)。人工序列例如包括天然存在的序列的突变体或者从头设计的序列。

本文所用术语“在严格条件下杂交”是指在 45°C, 在 6×氯化钠/柠檬酸钠 (SSC) 中杂交, 随后在 65°C, 在 0.2×SSC, 0.1% SDS 中洗涤两次的条件。

术语“结合优先性”是指多肽相对于另一个位点而选择一个核酸结合位点的判别力。例如在本文所述体内或体外分析中, 当多肽相对于两个不同的核酸结合位点而数量有限时, 更多量的多肽将结合相对于另一位点而优选的位点。

本文所用术语“简并寡核苷酸”是指这样的两种寡核苷酸: (a) 一群不同的寡核苷酸, 每个寡核苷酸均编码一特定的氨基酸序列, (b) 单一品种的寡核苷酸, 其能与一种以上的序列退火, 例如具有非天然核苷酸如肌苷的寡核苷酸。

本发明提供了许多益处。通过选择识别特定序列的 DNA 结合结构域的能力可以设计结合 DNA 上特异性位点的新多肽。因此, 本发明便于定制产生可以调节选择的靶表达的新多肽, 例如病原体需要的基因可被阻抑, 癌瘤生长需要的基因可被阻抑, 表达低下或编码不稳定蛋白质的基因可被激活及过表达等等。

锌指结构域的应用是特别有益的。首先, 锌指基序识别非常多样的 DNA 序列。其次, 天然存在的锌指蛋白的结构是模块化的。例如, 锌指蛋白 Zif268, 也称为 Egr-1, 由三个锌指结构域的串联阵列组成。图 1 是锌指结构域 Zif268 的 X 线晶体结构, 由与 DNA 复合的三个指

状结构组成 (Pavletich 和 Pabo, (1991) 科学 252:809—817)。每个指状结构独立地与 3—4bp 的 DNA 识别位点接触。通过在相同多肽链中具有多个锌指模块的协同效应达到高亲和性结合。

最终的目标是获得在细胞中发挥作用的 DNA 结合多肽。有利地, 体内选择方法鉴别了 up-front DNA 结合多肽, 其在细胞内的特异 DNA 位点发挥作用。与在细胞尤其真核细胞中识别相关的因子, 与在体外选择期间存在的因子有许多不同。例如在真核细胞核中, 多肽必须与无数的其它核蛋白竞争特异性核酸结合位点。核小体或另外的染色质蛋白可占据, 封闭或竞争结合位点。即使不结合, 细胞中的核酸构象也发生卷曲, 超螺旋, 扭转和解旋。所述多肽自身暴露于蛋白酶及陪伴分子等其它因子。另外, 所述多肽面临可能的结合位点的一个完整基因组, 因此对希望位点必须赋予高特异性以在选择过程中存留。与体内选择相反, 体外选择可以选择最高亲和性的结合子而不是最高特异性的结合子。

报道基因在表明表达的多肽嵌合体的结合能力中的应用不仅是有效及便利的, 而且避免了需要解释一种复杂的相互作用密码, 该密码表明蛋白质—核酸界面及也影响结合界面的无数周围因素如周围残基和核苷酸的能量学原因 (Segal 等 (1999), 美国科学院院报 96:2758—2763)。

本发明有利于人基因组或任何其它基因组中存在的所有锌指结构域。由锌指结构域结构折叠而占据的序列空间的这种多样化, 在天然选择阶段可以具有另外的益处。另外, 通过利用宿主物种的结构域, 针对基因治疗应用的本发明所述方法工程化的 DNA 结合蛋白降低了被宿主免疫应答认为是外源的可能性。

通过本发明方法鉴别的 DNA 结合蛋白可用于许多应用中。例如, 所述 DNA 结合蛋白可用于改变内源基因在培养的细胞或在宿主机体内的细胞中的表达。所述 DNA 结合蛋白可用于改变细胞的表型, 例

如增强对条件（例如应激）的敏感性或抗性，改变增殖能力，改变致病性及改变产物产生（例如代谢物产生）。

在此所引用的所有专利，专利申请及参考文献均以其全文并入参考。下文件有附图详细描述了本发明的一或多个实施方案。本发明的其它特征，目的和优势通过以下详细描述，附图和权利要求书显而易见。

附图简述

图 1 是 Zif268 锌指蛋白的三维结构图，其由三个指状结构域组成并结合 DNA 序列 5'-GCG TGG GCG T-3' (SEQ ID NO:197)。黑色圆点表示锌离子的位置。

图 2 是 Zif268 的氨基酸残基与 DNA 碱基之间氢键相互作用的图示。沿着 α 螺旋的第一1, 2, 3 和 6 位的氨基酸残基与在特异位置的碱基相互作用。粗线表示理想的氢键，虚线表示潜在的氢键。

图 3 是识别密码子表，其概括了 DNA 碱基与沿着锌指结构域 α 螺旋的第一1, 2, 3 和 6 位的氨基酸残基之间的相互作用。

图 4 是氨基酸残基及其相应的 3 碱基三联体的位置示意图。粗线表示观测到的主要相互作用，虚线表示辅助相互作用。

图 5 是文中所述体内选择系统的原理示意图。在各种锌指突变体中，锌指结构域 A 识别靶序列（称为 XXX X）并激活 *HIS3* 报道基因的转录。结果，酵母菌落在没有组氨酸的培养基上生长。相反，锌指结构域 B 不识别靶序列并因此报道基因还处于阻抑状态。结果，在没有组氨酸的培养基上无菌落生长。AD 表示转录激活结构域。

图 6 是在 HIV-1 的长末端重复 (LTR) 和 CCR5 (编码 HIV-1 共同受体的人基因) 的启动子区域中发现的一系列 10bp 序列（分别为 SEQ ID NO:1-5）。下划线处表示在本选择中使用的 4bp 靶序列。

图 7 是与报道基因连接的结合位点的碱基序列（分别为 SEQ ID NO:1-7）。每个结合位点由 4 个复合结合序列的一个串联阵列组成。

每个复合结合序列通过将由 Zif268 的第 1 指(finger 1)和第 2 指(finger 2) 识别的截短的结合序列 5'-GG GCG-3'与 4bp 靶序列连接而构建。

图 8 是 pPCFMS-Zif268 的示意图, 这是一种可用于构建杂合质粒文库 (SEQ ID NO:18 和 19) 的质粒。

图 9 是插入 pPCFMS-Zif 中的编码 Zif268 锌指蛋白的基因的碱基序列及相应的翻译的氨基酸序列 (分别为 SEQ ID NO:20 和 21)。下划线处表示限制酶识别的位点。

图 10 是培养平板的照片, 所述平板具有使用通过体内选择系统选择的锌指蛋白再转化和交叉转化获得的酵母细胞。

图 11A 是 P3 的多接头区域的核苷酸序列 (SEQ ID NO:251)。这个区域外侧的序列与亲代载体 pcDNA3 (Invitrogen) 的序列相同。每个酶位点均用斜体字标示, 下划线处是 HA 标记。起始和终止密码子均用黑体字母标示。图中还标明了核定位信号 (NLS)。

图 11B 是锌指蛋白文库构建的一个举例方法示意图。

图 12 是报道基因构建体及其在起始子区域内的序列节段的示意图。5XGal4, TATA 和 Inr 分别表示 5 个 GAL4 结合位点, TATA 框和转录起始子。NNNNNNNNN 表示特异性 ZFP 的关联结合位点。这个位置根据转录起始点 (+1) 编号, 相同核苷酸用 “-” 表示。“>” 表示相应核苷酸的缺失。

发明详述

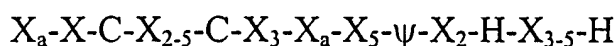
本发明涉及一种新的筛选方法以确定测试锌指结构域的核酸结合优先性。所述方法方便地适用于各种蛋白质结构域, 这些结构域的各种来源及许多文库设计, 报道基因, 及选择和筛选系统。筛选方法可作为高通量平台实施。得自筛选方法的信息易应用于设计人工核酸结合蛋白典型是 DNA 结合蛋白的方法中, 但在一些情况中也用于设计 RNA 结合蛋白或者甚至与其它蛋白质相互作用的蛋白质。所述设计方法根据测试锌指结构域的结合优先性指导嵌合核酸结合蛋白的

模块装配。所涉及的蛋白质可进一步优化或根据筛选方法而变化。

DNA 结合结构域

本发明利用具有不同结合特异性的核酸结合结构域的集合。已知与核酸的结合具有高亲和性和高特异性的各种蛋白质结构。这些结构重复用于无数的不同蛋白质中以特异性控制核酸功能（识别双链 DNA 的结构基序的综述见例如 Pabo 和 Sauer（1992），生物化学综述年刊 61:1053—95；Patikoglou 和 Burley（1997）生物物理学生物分子学结构综述年刊 26:289—325；Nelson（1995）Curr Opin Genet Dev 5:180—9）。核酸结合结构域的几个非限制性实例包括：

锌指：锌指是大约 30 个氨基酸残基的小多肽结构域，其中有 4 个残基，半胱氨酸或组氨酸，合适地隔开由此能配位锌离子（图 1；见例如 Klug 和 Rhodes，生物化学科学趋势 12:464—469（1987）；Evans 和 Hollenberg（1988），细胞 52:1—3；Payre 和 Vincent（1988），FEBS 通讯 234:245—250；Miller 等（1985），EMBO 杂志 4:1609—1614；Berg（1988），美国科学院院报 85:99—102；Rosenfeld 和 Margalit（1993），生物分子学结构动力学杂志 11:557—570）。因此，锌指结构域可以根据配位锌离子的残基的身份进行分类，例如 Cys₂-His₂ 类，Cys₂-Cys₂ 类，Cys₂-CysHis 类等。Cys₂-His₂ 锌指的锌离子配位残基典型如下间隔：



其中 ψ (psi) 是一个疏水性残基（Wolfe 等（1999），生物物理学生物分子学结构综述年刊 3:183—212）（SEQ ID NO:76），“X”代表任何氨基酸， X_a 是苯丙氨酸或酪氨酸，下标数表示氨基酸的数目，有两个连字符连接的数字下标表示间隔氨基酸的典型范围。典型地，间隔氨基酸折叠形成反平行 β 折叠，其针对 α 螺旋反方向卷曲，所述反平行 β 折叠可以是短的，非理想的或不存在的。所述折叠调整锌离子配位侧链，因此它们呈适于配位锌离子的四面体构象。碱基接触残

基在锌指的 N 末端及在前面的环状区域内 (图 2)。

为方便起见, 锌指结构域的一级 DNA 接触残基基于以下实例编号为-1, 2, 3 和 6:

-1 1 2 3 4 5 6

$X_a-X-C-X_{2-5}-C-X_3-X_a-X-C-X-S-N-X_b-X-R-H-X_{3-5}-H$ (SEQ ID NO:68)

如上述实例所指出的, DNA 接触残基是 Cys(C), Ser(S), Asn(N) 和 Arg(R)。以上基序可缩写为 CSNR。如本文所用, 这种缩写是一种速记形式, 是指从第一个半胱氨酸之前的第二个残基 (X_a , 上述 SEQ ID NO:68 的起始残基) 至最后的金属螯合组氨酸 (SEQ ID NO:68 的最后的残基) 的一个特定的多肽序列。在两个不同序列具有相同基序的情况下, 可以使用数字标明每个序列 (例如 CSNR1, CSNR2)。在一些更明确的情况中, 四个字母缩写一般是指所述基序。

锌指 DNA 结合蛋白可由三或多个锌指结构域的串联阵列组成。

锌指结构域 (或 ZFD) 是最普遍的真核细胞 DNA 结合基序之一, 在从酵母至高等植物至人的多个物种中发现。据估计只在人基因组中就有至少几千个锌指结构域, 可能至少 4500 个。锌指结构域可以在锌指蛋白中鉴别或从中分离。锌指蛋白的非限制性实例包括 CF2-II; Kruppel; WT1; 碱性核蛋白; BCL-6/LAZ-3; 红细胞类 Kruppel 转录因子; 转录因子 Sp1, Sp2, Sp3 及 Sp4; 转录阻抑因子 YY1; EGR1/Krox24; EGR2/Krox20; EGR3/Pilot; EGR4/AT133; Evi-1; GLI1; GLI2; GLI3; HIV-EP1/ZNF40; HIV-EP2; KR1; ZfX; ZfY 和 ZNF7。

可以使用以下所述计算方法鉴别在测序的基因组或核酸数据库中编码的所有锌指结构域。可以利用任何这种锌指结构域。另外, 使用计算方法 (例如 Dahiyat 和 Mayo (1997), 科学 278:82-7) 已经设计了人工锌指结构域。

尽管许多锌指结构域与 DNA 位点结合，但许多锌指结构域可结合其它配体，例如 RNA 位点及其它蛋白质。在一些应用中，嵌合的锌指结构域蛋白工程化为结合非 DNA 配体，例如靶蛋白或靶 RNA 位点。所述靶 RNA 位点可以是 ncRNA 上的位点，例如天然存在的 ncRNA 上的位点。

同源域：同源域是简单的真核细胞结构域，由接触 DNA 小沟的 N 末端臂后接接触大沟的三个 α 螺旋组成（见例如 Laughon (1991)，生物化学 30:11357—67）。第三个 α 螺旋位于大沟内，并含有关键的 DNA 接触侧链。同源域在转角具有特征性的高保守的基序，形成第三个 α 螺旋。所述基序包括一个不变的色氨酸，其卷曲在该结构域的疏水核心中。这个基序在 Prosite 数据库（见 Falquet 等 (2002)，核酸研究 30:235 — 238）中表示为 PDOC00027 ([L/I/V/M/F/Y/G]-[A/S/L/V/R]-X(2)-[L/I/V/M/S/T/A/C/N]-X-[L/I/V/M]-X(4)-[L/I/V]-[R/K/N/Q/E/S/T/A/I/Y]-[L/I/V/F/S/T/N/K/H]-W-[F/Y/V/C]-X-[N/D/Q/T/A/H]-X(5)-[R/K/N/A/I/M/W]; SEQ ID NO:77)。同源域通常在转录因子中发现，其确定细胞身份并在机体发育期间提供位置信息。这种典型的同源域可在基因组中成群发现，由此同源域在群体中的顺序几乎相当于其沿着体轴的表达模式。同源域可以通过与一个同源域例如 Hox-1 进行序列对比而鉴别，或者通过与同源域分布图或同源域隐匿 Markov 模型（HMM，见下文）例如 Pfam 数据库的 PF00046 或者 SMART 数据库的 HOX 进行序列对比而鉴别，或者通过上述 Prosite 基序 PDOC00027 鉴别。

螺旋—转角—螺旋蛋白：这种 DNA 结合基序在许多原核细胞转录因子中是很普遍的。有许多亚家族，例如 LacI 家族，AraC 家族等。名称中的两个螺旋是指第一个 α 螺旋，其针对第二个螺旋卷曲并使之位于 DNA 大沟内。这些结构域可以通过与 HMM 进行序列对比而鉴别，所述 HMM 例如是得自 SMART 数据库的 HTH-ARAC，

HTH-ARSR, HTH-ASNC, HTH-CRP, HTH-DEOR, HTH-DTXR, HTH-GNTR, HTH-ICLR, HTH-LACI, HTH-LUXR, HTH-MARR, HTH-MERR 及 HTH-XRE 分布图。

螺旋一环一螺旋蛋白：这个 DNA 结合结构域普遍见于同源二聚体和异源二聚体转录因子中，所述转录因子如 MyoD, fos, jun, E11 和肌细胞生成素。所述结构域由二聚体组成，每个单体提供两个 α 螺旋以及介于其间的环。所述结构域可以通过与一种 HMM 进行序列对比而鉴别，所述 HMM 例如是得自 SMART 数据库的 HLH 分布图。尽管螺旋一环一螺旋蛋白是典型的二聚体，但单体形式可通过在两个亚单位之间工程化一个多肽接头而构建，由此一个单一的开放读框既编码这两个亚单位又编码接头。

DNA 结合结构域的鉴别

可使用各种各样的方法鉴别结构域。

计算方法：通过本发明所述方法分离的 DNA 结合结构域的氨基酸序列可以与已知序列的数据库对比，所述数据库例如是蛋白质序列的注解数据库或者包括核酸结合结构域条目的注解数据库。在另一个实施方案中，未鉴定的序列例如未注解的基因组，EST 或全长 cDNA 序列的数据库；鉴定的序列的数据库例如 SwissProt 或 PDB；及结构域的数据库例如 Pfam, ProDom (Corpet 等 (2000), 核酸研究 28:267—269), 及 SMART (简便模块构筑研究工具, Letunic 等 (2002) 核酸研究 30, 242—244), 可以提供核酸结合结构域序列的来源。核酸序列数据库可以在全部 6 个读框中翻译，以与查询氨基酸序列对比。标记为编码候选核酸结合结构域的核酸序列可以从合适的核酸来源中扩增，例如从基因组 DNA 或细胞 RNA 中扩增。这种核酸序列可以克隆入表达载体中。基于计算机的结构域鉴别程序可以与寡核苷酸合成仪及自动化系统接合，在高通量平台中生产编码所述结构域的核酸。编码候选结构域的克隆的核酸也可以贮存于宿主表达载体中并易

于穿梭进一种表达载体中，例如进入具有 Zif268 第 1 指和第 2 指的翻译融合载体中，这通过限制酶介导的亚克隆或者通过位点特异性重组酶介导的亚克隆进行（见美国专利 No.5888732）。所述高通量平台可用于产生多个含有编码不同候选核酸结合结构域的核酸的微滴定平板。

本领域熟知从起始序列或分布图中鉴别结构域的详细方法。见例如 Prosite (Hofmann 等 (1999), 核酸研究 27:215—219), FASTA, BLAST (Altschul 等 (1990), 分子生物学杂志 215:403—10) 等。可以进行简单字符串检索以发现与查询序列或查询分布图具有相同性的氨基酸序列，例如使用 Perl 扫描文本文件而进行。如此鉴别的序列与起始的输入序列可以有至少大约 30%, 40%, 50%, 60%, 70%, 80%, 90%或更高的相同性。

与查询结构域相似的结构域可以从公共数据库中鉴别，例如使用 Altschul 等所述 XBLAST 程序（版本 2.0）（分子生物学杂志 215:403—10）。例如，可使用如下 XBLAST 参数进行 BLAST 蛋白检索：score = 50, wordlength = 3。在查询或检索的序列中可导入缺口，如 Altschul 等 (1997), 核酸研究 25(17):3389—3402 所述。XBLAST 和 Gapped BLAST 程序的默认参数得自国际生物工程信息中心 (NCBI), 国立卫生研究院, Bethesda, MD。

可以使用 Prosite 分布图 PS00028 和 PS50157 鉴别锌指结构域。在具有 80000 个蛋白质序列的 SWISSPROT 中，这些分布图分别检测到 3189 和 2316 个锌指结构域。分布图可通过各种不同的方法从相关蛋白的多个序列对比中构建。Gribskov 及同事 (Gribskov 等, (1990) 酶学方法 183:146—159) 利用符号对比表将具有残基分布频率的多个序列对比转变为针对每个位置的加权。见例如 PROSITE 数据库及 Luethy 等 (1994), 蛋白质科学 3:139—1465 所述。

代表感兴趣的 DNA 结合结构域的 Hidden Markov 模型 (HMM's)

可以产生自或得自这种模型的数据库,例如 Pfam 数据库, release2.1。数据库可例如使用 HMM 利用默认参数进行检索,以发现额外的结构域(见例如 Bateman 等(2002), 核酸研究 30:276—280)。或者,用户可以优化这些参数。可以选择极限记录以过滤序列数据库,从而分值在极限之上的序列展示为候选结构域。关于 Pfam 数据库的描述可见于 Sonhammer 等,(1997) 蛋白质 28(3): 405—420, 关于 HMM 的详细描述见例如 Gribskov 等(1990), 酶学方法 183:146—159; Gribskov 等(1987), 美国科学院院报 84:4355—4358; Krogh 等,(1994) 分子生物学杂志 235:1501—1531; 及 Stultz 等(1993), 蛋白质科学 2:305—314。

HMM 的 SMART 数据库(简便模块构筑研究工具, Schultz 等(1998) 美国科学院院报 95:5857 及 Schultz 等,(2002) 核酸研究 28:231), 提供了锌指结构域的目录(ZnF_C2H2; ZnF_C2C2; ZnF_2HC; ZnF_3H1; ZnF_4; ZnF_HCC; ZnF_ATA; 及 ZnF_FX), 所述锌指结构域是通过 HMMer2 检索程序的 Hidden Markov 模型通过作出分布图而鉴别的(Durbin 等,(1998) 生物学序列分析: 蛋白质和核酸的概率模型, 剑桥大学出版社)。

基于杂交的方法: 可以分析编码各种形式 DNA 结合结构域的核酸集合以作出编码保守的氨基和羧基末端边界序列的序列分布图。可以设计简并寡核苷酸以与编码这种保守的边界序列的序列杂交。另外, 这种简并寡核苷酸的效力可以通过将其组成与已知基因组序列中的可能退火位点的频率进行对比而评估。可以使用多轮设计以使简并寡核苷酸最佳化。例如, 一些已知 Cys₂-His₂ 锌指结构域的对比示出在天然序列中相邻指之间的接头区域中的一个公有序列(Agata 等(1998), 基因 213:55—64)。这种简并寡核苷酸用于扩增众多的 DNA 集合结构域。将扩增的结构域作为测试锌指结构域插入杂合核酸中, 随后通过本发明所述方法分析与靶位点的结合。

核酸结合结构域的集合

所述方法可以筛选编码 DNA 结合结构域的核酸集合(例如质粒,噬菌粒或噬菌体文库形式)的核酸结合性质。所述集合可编码不同的 DNA 结合结构域,甚至不同折叠结构的结构域。在一个例子中,所述集合编码单一折叠结构的结构域如锌指结构域。尽管以下方法是针对锌指结构域进行描述的,但本领域技术人员可以将所述方法适用于其它类型的核酸结合结构域。

突变的结构域:在另一个实例中,所述集合由编码从简并的模式文库中装配的结构域的核酸组成。例如,在锌指情况中,已知锌指的序列对比可以用于鉴别在每个位置的最佳氨基酸。或者,结构研究和诱变实验可用于确定在每个位置的氨基酸的优选性质。任何核酸结合结构域均可用作结构支架以导入突变。特别地,可以定向在与核酸结合界面紧密接近或与其相邻的位置进行诱变。通过使用模式化简并文库,突变的测试锌指结构域可以被限制在可能的氨基酸亚群的任何突变位置。简并密码子系列可用于在每个位置编码分布图。例如,可获得只编码疏水性残基,脂族残基或亲水性残基的密码子系列。可以针对编码折叠的多肽的全长克隆选择文库。Cho 等((2000),分子生物学杂志 297(2): 309-19)提供了一种使用简并的寡核苷酸生产这种简并文库的方法,还提供了一种选择编码全长多肽的文库核酸的方法。这种核酸可被插入表达质粒中,针对所述的选择方法使用常规限制酶切割位点或转座酶或重组酶识别位点进行这种插入。

合适的密码子的选择及在给定位置的每个核苷酸的相关性质,可以通过简便检测遗传密码表而确定,或者通过计算公式确定。例如,Cho 等(如前)描述了一个计算机程序,其接受希望的蛋白质序列分布图并输出一个编码分布图序列的优选的寡核苷酸设计。例如,所述设计可包括寡核苷酸群的简并位置。

分离结构域的天然的所有组分:可以从真核生物体如人的基因组

DNA 或 cDNA 中构建结构域文库。为此有许多方法可利用。例如，可以使用上述氨基酸序列的计算机检索以鉴别结构域。可分离编码每个结构域的核酸并插入适于在细胞中表达的载体中，例如含有启动子，激活结构域和选择标记的载体。在另一个实施例中，与编码保守的基序的序列杂交的简并寡核苷酸用于例如通过 PCR 扩增含有所述基序的大量相关结构域。例如类 Kruppel 的 Cys₂His₂ 锌指可以通过 Agata 等 (1998) 基因 213:55—64 所述方法扩增。这个方法还保留了天然存在的锌指结构域接头肽序列，例如具有以下模式的序列：Thr-Gly-(Glu/Gln)-(Lys/Arg)-Pro-(Tyr/Phe)(SEQ ID NO:78)。另外，筛选限于感兴趣的结构域集合与筛选未选择的基因组或 cDNA 序列不同，前者明显降低文库的复杂性并降低由于完全筛选大文库的固有难度而丢失合意的序列的可能性。

人基因组含有许多锌指结构域，许多是未鉴定及未鉴别的。据估计有上千个编码具有锌指结构域的蛋白质的基因 (Pellegrino 和 Berg (1991), 美国科学院院报 88:671—675)。这些人锌指结构域代表不同结构域的广泛集合，从中可以构建新的 DNA 结合蛋白。如果每个锌指结构域识别一个独特的 3—4bp 的序列，则结合每个可能的 3—4bp 序列需要的结构域的总数仅为 64—256 ($4^3—4^4$) 个。人基因组的天然所有组分含有足够数目的独特锌指结构域以跨越所有可能的识别位点是可能的。这些锌指结构域是构建人工嵌合 DNA 结合蛋白的重要资源。天然存在的锌指结构域，与衍生自人基因组的人工突变体不同，其已经在天然选择性压力下进化并因此针对结合特异性 DNA 序列及体内功能是天然最佳化的。

人锌指结构域当导入人体时，例如在基因治疗应用中，不太可能诱导免疫应答。

具有特异性 DNA 结合性质的锌指结构域的体内选择

具有希望的 DNA 识别性质的锌指结构域，可以使用以下所述体

内筛选系统进行鉴别。将感兴趣的一个复合 (composite) 结合位点插入报道基因的上游, 由此转录激活结构域对所述复合结合位点的募集导致报道基因转录水平高于给定水平。构建一个编码杂合蛋白的表达质粒, 所述杂合蛋白由与固定的 DNA 结合结构域融合的一个测试锌指结构域和一个转录激活结构域组成。

所述复合结合位点包括至少两个元件, 即一个募集位点和一个靶位点。将该系统工程化, 由此固定的 DNA 结合结构域识别所述募集位点。然而, 固定的 DNA 结合结构域与募集位点的结合亲和性是这样的, 即在体内其独自不足以转录激活报道基因。这可以通过对照实验证实。

例如, 当在细胞中表达时, 固定的 DNA 结合结构域 (不存在测试锌指结构域, 或存在已知是无功能的或者其已知 DNA 接触残基已经由另一种氨基酸如丙氨酸置换的测试锌指结构域) 应不能激活报道基因的转录。一些渗漏或低水平激活是可耐受的, 因为所述系统可以通过其它方式敏化 (例如通过使用报道基因的竞争性抑制剂)。所述固定的 DNA 结合结构域预期不稳定结合所述募集位点。例如, 固定的 DNA 结合结构域可以结合所述募集位点, 解离常数 (K_d) 为至少大约 0.1nM, 1nM, 1 μ M, 10 μ M, 100 μ M 或更高。(另外, 所述 K_d 可低于 100 μ M 或 10 μ M)。DNA 结合结构域与靶位点的 K_d 可以在体外测定, 通过无测试锌指结构域或无具有特异于第二个靶位点的测试锌指结构域的情况下, 通过电泳迁移率变动分析 (EMSA) 测定。

因此, 识别靶位点例如所述复合结合位点的可变位点的功能性测试锌指结构域的附着, 是杂合蛋白稳定结合细胞中的复合结合位点所必需的, 从而激活报道基因。测试锌指结构域与靶位点的结合优先性导致报道基因表达相对于给定水平提高。例如, 通过将观测到的水平除以给定水平获得的报道基因表达提高倍数可至少为大约 2, 4, 8, 20, 50, 100, 1000 倍或更高。当测试锌指结构域识别靶位点时, 包

含 DNA 结合结构域和测试锌指结构域的转录因子的 K_d 降低, 例如相对于没有对靶位点具有特异性的测试锌指结构域的转录因子的 K_d 。例如, 与其特异性靶位点复合的转录因子的解离常数 (K_d) 可以等于或低于大约 50nM, 10nM, 1nM, 0.1nM, 0.01nM 或更低。所述 K_d 可以在体外通过 EMSA 测定。

关于 DNA 结合特异性可以灵敏化而精确地通过测定测试锌指结构域增强固定的 DNA 结合结构域的体内结合亲和性的能力而分析地发现, 使得可以从人基因组中快速分离并鉴定新的锌指结构域。

固定的 DNA 结合结构域包括分离自天然存在的 DNA 结合蛋白的模块结构域, 所述蛋白例如是具有多个结构域的或者是一种寡聚物的天然存在的 DNA 结合蛋白。例如, 包括两个已知锌指例如 Zif268 的第 1 指和第 2 指的氨基酸序列, 可以用作固定的 DNA 结合结构域。技术人员能从无数的核酸结合结构域中鉴别一种适于所述系统的固定的 DNA 结合结构域, 所述无数的核酸结合结构域例如是在此所述结构域家族如同源域, 螺旋—转角—螺旋结构域, 或者螺旋—环—螺旋结构域, 或者本领域充分鉴定的核酸结合结构域。还必需适当选择由固定的 DNA 结合结构域识别的募集位点。募集位点可以是从中获得固定的 DNA 结合结构域的天然存在的 DNA 结合蛋白的天然结合位点内的一个亚位点。如果必需, 可以将突变导入固定的结构域或募集位点中以敏化所述系统。

适于体内筛选系统的细胞包括真核细胞和原核细胞。真核细胞例如包括酵母细胞例如 *Saccharomyces cerevisiae*, *Saccharomyces pombe* 及 *Pichia pastoris* 细胞。

对使用 *Saccharomyces cerevisiae* 的酵母单杂交系统加以修改, 以使用前述筛选系统选择锌指结构域。首先, 制备编码 HIS3 报道基因的报道基因质粒。将预先测定的 4bp 靶 DNA 序列与截短的结合序列连接, 以提供所述 DNA 结合结构域的复合结合序列, 每个复合结合

序列在单独的质粒上可操纵地与报道基因连接。

所述杂合核酸序列编码与 DNA 结合结构域连接的一种转录激活结构域，所述 DNA 结合结构域包含一个截短的 DNA 结合结构域和一个锌指结构域。

在此所用的结合位点非必需是连续的，尽管通常使用连续的位点。可以使用在核酸结合结构域之间的柔性和/或可伸展接头构建识别非连续位点的蛋白质。

根据本发明的一方面，由 Zif268 的第 1 指和第 2 指组成的及无第 3 指的多肽可用作固定的 DNA 结合结构域。（在 Zif268 的三个锌指结构域之中，第 1 指是位于 N 末端的锌指结构域，第 2 指是在中间的锌指结构域，第 3 指是在 C 末端的锌指结构域）。或者，一结合位点已被鉴定的任两个锌指结构域可用作固定的 DNA 结合结构域。在此揭示了许多新的例子。

其它有用的固定的 DNA 结合结构域可衍生自其它锌指蛋白，如 Sp1, CF2-II, YY1, Kruppel, WT1, Egr2, 或 POU 结构域蛋白，如 Oct1, Oct2 和 Pit1。这些只是例证而无限制本发明之意。

根据本发明的一个特定的实施例，通过在最佳 Zif268 识别序列（5'-GCG TGG GCG-3'）的 5'末端缺失 4bp 而产生的 5'-GGGCG-3' 碱基序列，可以用作募集位点。3-4bp 的任何靶序列可以与这个募集位点连接，以产生复合的结合序列。

激活结构域：可用于本发明中的转录激活结构域包括但非限于酵母的 Gal4 激活结构域及单纯疱疹病毒的 VP16 结构域。在细菌中，激活结构域的功能可通过融合一个能募集野生型 RNA 聚合酶 α 亚单位 C 末端结构域或者突变的 α 亚单位 C 末端结构域而模仿，例如一个 C 末端结构域与蛋白质相互作用结构域融合。

阻抑结构域：如果需要，阻抑结构域而不是激活结构域可以与 DNA 结合结构域融合。真核细胞阻抑结构域例如包括 ORANGE，

groucho 和 WRPW (Dawson 等 (1995), 分子细胞生物学 15:6923—31)。当使用阻抑结构域时, 可以使用毒性报道基因和/或不可选择的标记筛选降低的表达水平。

其它功能性结构域: 蛋白质转导结构域可以与 DNA 结合结构域融合, 例如与嵌合的锌指蛋白的 DNA 结合结构域融合。蛋白质转导结构域导致转导结构域及附着的多肽吸收入细胞中。蛋白质转导结构域的一个实例是 HIV tat 蛋白。

报道基因: 报道基因可以是一个选择标记, 例如授予药物抗性的基因或者营养缺陷型标记。药物抗性基因包括例如 *S.cerevisiae* 环己酰亚胺抗性 (CYH) 基因, *S.cerevisiae* 刀豆氨酸抗性 (CAN1) 基因及潮霉素抗性基因。*S.cerevisiae* 营养缺陷型标记包括 URA3, HIS3, LEU2, ADE2 及 TRP1 基因。当营养缺陷型标记是报道基因时, 利用没有营养缺陷型基因的功能性拷贝并因此无产生特定代谢物的能力的细胞。编码结合靶位点的测试锌指结构域的构建体的选择通过无代谢物的培养基中维持所述细胞而实现。例如, HIS3 基因可用作选择标记与 his3⁻酵母株组合。在导入编码杂合转录因子的构建体之后, 将细胞在没有组氨酸的情况下生长。本领域技术人员还熟知用于哺乳动物细胞中的选择标记, 如胸苷激酶, 新霉素抗性 & HPRT。

或者, 所述报道基因编码一种蛋白质, 其存在可易于检测和/或定量。所述报道基因例如包括 lacZ, 氯霉素乙酰转移酶 (CAT), 萤光素酶, 绿色荧光蛋白 (GFP), β -葡糖醛酸糖苷酶 (GUS), 蓝色荧光蛋白 (BFP), 及 GFP 的衍生物, 例如具有改变的或增强的荧光性质 (Clontech Laboratories, Inc., CA)。表达 lacZ 的细胞菌落可通过将所述菌落在含有生色底物 X-gal 的平板上生长而便利检测。GFP 表达可通过监测在激发时的荧光发射而检测。各个 GFP 表达细胞可以使用荧光激活的细胞分选 (FACS) 鉴别并分离。

所述系统可以用两个报道基因构建, 例如可选择的报道基因和不

可选择报道基因。所述选择标记便于快速鉴别感兴趣的结构域，在合适的生长条件下，只有携带感兴趣的结构域的细胞生长。不可选择的报道基因提供了一种确证方法，例如区别假阳性，及一种定量结合程度的方法。两个报道基因可以整合在基因组中单独的位置，串联整合在基因组中，包含于相同的染色体外元件（例如质粒）上或者包含于单独的染色体外元件上。

图 5 例证了修改的单杂交系统用于选择希望的锌指结构域的原理。所述杂合转录因子的 DNA 结合结构域由 (a) 由 Zif268 的第 1 指和第 2 指组成的一个截短的 DNA 结合结构域，(b) 锌指结构域 A 或 B 组成。位于报道基因的启动子区域的结合位点的碱基序列是一个复合的结合序列 (5'-XXXXGGGCG-3')，其由一个 4bp 的靶序列（核苷酸 1-4，5'-XXXX-3'）和一个截短的结合序列（核苷酸 5-9，5'-GGGCG-3'）组成。

如果杂合转录因子中测试锌指结构域（图 5 中 A）识别靶序列，则所述杂合转录因子可稳定结合复合的结合序列。这种稳定结合通过杂合转录因子的激活结构域（图 5 的 AD）的作用导致报道基因表达。结果，当 HIS3 用作报道基因时，转化的酵母在无组氨酸的培养基中生长。或者，当 lacZ 用作报道基因时，转化的酵母在含有 X-gal 的培养基中以蓝色菌落形式生长，所述 X-gal 是 lacZ 蛋白的底物。然而，如果杂合转录因子的锌指结构域（图 5 中 B）不识别靶序列，则报道基因的表达不被诱导。结果，转化的酵母不能在没有组氨酸的培养基中生长（当 HIS3 用作报道基因时），或者在含有 X-gal 的培养基中以白色菌落形式生长（当 lacZ 用作报道基因时）。

使用这种修改的单杂交系统的选择方法是有益的，因为通过这种方法选择的锌指结构域表明在细胞环境中起作用。因此，推测所述结构域能折叠，进入细胞核并抵抗胞内蛋白酶及其它有潜在损害作用的胞内因子的作用。另外，本发明揭示的修改的单杂交系统可以快速而

便利地分离希望的锌指结构域。所述修改的单杂交系统仅需要一轮酵母细胞转化就可以分离希望的锌指结构域。

可以利用所述选择方法从基因组例如植物或动物物种（例如哺乳动物，例如人）基因组中鉴别锌指结构域。还可利用所述方法从突变的锌指结构域的文库中鉴别锌指结构域，所述突变的锌指结构域文库例如是通过随机诱变制备的。另外，可以组合使用这两种方法。例如，如果针对特定的 3 或 4bp DNA 序列的锌指结构域不能从人基因组中分离，可对通过随机或定向诱变制备的锌指结构域的文库筛选这种结构域。

尽管修改的单杂交系统在酵母中是选择识别并结合给定靶序列的锌指结构域的优选方式，但本领域技术人员显然可以使用除了酵母单杂交选择之外的系统。例如，可使用噬菌体展示选择筛选衍生自真核生物体的天然存在的锌指结构域文库。

本发明涵盖了单杂交方法在各种培养细胞中的应用。例如，与靶序列可操纵连接的报道基因可导入培养的原核或动物或植物细胞中，然后培养的细胞可用编码锌指结构域文库的质粒，噬菌体或病毒转化。然后从报道基因被激活的分离的细胞中可获得希望的锌指结构域识别的靶序列。

以下揭示的实施例表明所述方法可鉴别结合感兴趣的位点的锌指结构域。制备在第 3 指具有各种锌指结构域的杂合转录因子的文库，从所述文库选择的新的锌指结构域（例如 HSNK, QSTV 和 VSTR 锌指结构域；见下文）中，无一天然位于其相应亲代锌指蛋白的 C 末端。这清楚表明锌指结构域是模块化的而且新的 DNA 结合结构域可以通过混合并匹配合适的锌指结构域而构建。

通过本发明方法选择的锌指结构域可用作构筑模块以通过合适的重排和重组产生新的 DNA 结合蛋白。例如，可如下构建识别人 CCR5 的启动子区域的新 DNA 结合蛋白，其是 HIV-1 的共同受体。

人 CCR5 的启动子区域含有以下 10bp 序列：5'-AGG GTG GAG T-3' (SEQ ID NO:4)。使用本发明揭示的修改的单杂交系统，可以分离三个锌指结构域，每个结构域特异性识别以下 4bp 靶序列之一：5'-AGGG-3'；5'-GTGG-3'和 5'-GAGT-3'。这些靶序列是 CCR5 靶序列的重叠 4bp 序列节段。这三个锌指结构域可与合适的接头连接并附着于一个调节结构域如 VP16 结构域和 GA14 结构域或者阻抑结构域如 KRAB 结构域，以产生特异性结合 CCR5 启动子的新的转录因子。可以设计相似的锌指以识别以下序列：

HIV-1 LTR(-124/-115): 5'-GAC ATC GAG C-3' (SEQ ID NO:1)

HIV-1 LTR(-23/-14): 5'-GCA GCT GCT T-3' (SEQ ID NO:2)

HIV-1 LTR(-95/-86): 5'-GCT GGG GAC T-3' (SEQ ID NO:3)

人 CCR5(-70/-79): 5'-AGG GTG GAG T-3' (SEQ ID NO:4)

人 CCR5(+7/+16): 5'-GCT GAG ACA T-3' (SEQ ID NO:5)

这些锌指结构域可用于治疗中以有助于预防 HIV-1 增殖。

高通量筛选

以下方法可快速测定集合中每个结构域与多个可能的 DNA 结合位点或者甚至所有可能的 DNA 结合位点的相对体内结合亲和性。产生编码核酸结合结构域的核酸大集合。每个核酸结合结构域在杂合核酸构建体中编码为测试锌指结构域，并在一个交配型酵母株中表达。因此产生第一系列的酵母株，其表达所有可利用的或希望的结构域。在相对应交配型中构建含有报道基因构建体的第二系列酵母株，所述报道基因构建体针对具有报道基因构建体中结构域的推定的靶位点。该方法需要进行许多或全部可能的成对交配，以产生融合细胞的矩阵，每个细胞均具有不同的测试锌指结构域和不同的靶位点报道基因构建体。分析每个融合细胞的报道基因表达情况。该方法因此可快速及容易地确定测试结构域的结合优先性。

结构域集合例如通过检索基因组数据库中符合给定分布图的推

定结构域而鉴别。所述集合可包括例如 10—20 个结构域，或者所有鉴别的结构域，可能有上千个或更多。编码从所述数据库中鉴别的结构域的核酸从头合成或者使用寡核苷酸从基因组 DNA 样品中扩增。设计这种合成的寡核苷酸的指导和自动化方法是本领域常规应用的。编码另外的结构域的核酸可以相似合成或者用简并引物扩增。将编码所述集合的结构域的核酸克隆入上述酵母表达质粒中，由此产生所述结构域和 Zif268 的前两个锌指及转录激活结构域的融合蛋白。在微滴定平板中进行扩增和克隆步骤，以克隆编码多个结构域的核酸。

或者，可以使用重组克隆方法将编码所述结构域的多个扩增的核酸快速插入酵母表达载体中。这种方法见美国专利 No.5888732 及“Gateway”指导（Life Technologies-Invitrogen, CA, 美国）所描述，需要在扩增引物的末端包括位点特异性重组酶的特定位点。所述表达载体在用于插入编码所述结构域的扩增核酸的位置含有一个或多个额外的位点。将这些位点设计为没有终止密码子。在重组反应中加入扩增产物，表达载体和位点特异性重组酶导致扩增的序列插入所述载体中。这种方法的其它特点例如在成功插入时置换毒性基因，使该方法是高效的及适于高通量克隆。

可以使用限制酶介导的和/或重组克隆以将编码每个鉴别的结构域的核酸插入表达载体中。所述载体可以在细菌中增殖，在标示的微滴定平板中冷冻，由此每个孔均含有一种细胞，所述细胞具有编码不同的独特的 DNA 结合结构域之一的核酸。

获得针对每个结构域的分离的质粒 DNA，并转化入酵母细胞中，例如 *Saccharomyces cerevisiae* MATa 细胞。由于表达载体含有一个可选择的标记，因此转化的细胞在选择所述标记的营养条件下在基本培养基中生长。这种细胞也可以冷冻并贮存，例如在微滴定平板中冷冻并贮存直至使用。

在例如 *Saccharomyces cerevisiae* MATa 细胞中构建第二系列酵母

株。这个系列酵母株含有各种不同的报道基因载体。然后将携带具有独特的 DNA 结合结构域的表达载体的每个酵母株与报道基因系列的每个酵母株交配。由于这两个酵母株来自相反的交配型并工程化为具有不同的辅源营养，因此可易于选择二倍体。这种二倍体既具有报道基因又具有表达质粒。所述细胞也可以在选择报道基因和表达质粒二者的营养条件下维持。Uetz 等 (2000)，自然 403: 623-7 通过产生这种酵母交配矩阵描述了所有酵母蛋白的一个完整的双杂交图。

报道基因表达可以以高容量形式检测，例如在微滴定平板中。例如，当使用 GFP 作为报道基因时，含有交配细胞矩阵的平板可以进行荧光扫描。

设计新的 DNA 结合蛋白

可以通过混合并匹配合适的锌指结构域合理地构建一种新的 DNA 结合蛋白以识别 9bp 或更长的 DNA 靶序列。锌指结构域的模块结构便于其重排以构建新的 DNA 结合蛋白。如图 1a 所示，天然存在的 Zif268 蛋白中的锌指结构域沿着 DNA 双螺旋串联排列。每个结构域独立识别一种不同的 3-4bp 的 DNA 节段。

锌指结构域的数据库：可利用上述单杂交选择系统针对每个可能的 3 或 4bp 结合位点或代表性数目的这种结合位点鉴别一或多个锌指结构域。这个方法的结果可积累为锌指结构域与其优选的 3 或 4bp 的结合位点之间关联级数。这种关联例如在表 3-6 中提供。

所述结果还可以存储于机器中作为数据库，例如相关数据库，电子表格或文本文件。这种数据库的每个记录将一个锌指结构域与表示所述结构域的一或多个优选的结合位点的序列的字符串相关联。所述数据库记录可包括结合每个位点的锌指结构域的相对亲和性指标。在一些实施方案中，所述数据库记录也可包括表明编码特定的锌指结构域的核酸的物理位置的信息。这种物理位置可例如是贮存于冰箱中的微滴定平板的特定孔。

所述数据库可被配置成可以例如使用 SQL 操作环境，脚本（scripting）语言（如 PERL 或 Microsoft Excel®macro），或者程序语言进行查询或过滤。这种数据库使用户可以鉴别识别特定 3 或 4bp 结合位点的一或多个锌指结构域。数据库及其它信息如可以存储于数据库服务器上的信息也可以配置成使用可由所述设备解读的命令及其它信号与每个设备沟通。所述系统基于计算机方面可以安装在数字式电子电路中，或计算机硬件，固件，软件中，或者在其组合中。本发明的设备例如数据库服务器，可以在确实收录在通过可编程处理器实现的机读存储设备中的计算机程序产品形式应用；所述方法的作用可以通过执行指令程序的可程序化处理器进行，以通过运行输入数据及产生输出数据而完成本发明功能。实现环境的一个非限制性实例包括计算机操作系统 Windowsv XP 或者 Windows NT4.0（Microsoft）或更好或者 Solaris 2.6 或更好（Sun Microsystems）的操作系统。

锌指结构域也可以在多个不同的融合蛋白质中测试以证实其特异性。另外，极少量结构域可利用的特定结合位点可以是另外的选择筛选的目标。这种选择的文库可以通过诱变结合相似但仍独特的位点的锌指结构域而制备。每个可能的结合位点的锌指结构域的完整矩阵是非必需的，因为所述结构域可以与靶结合位点交错（staggered）相关以最佳利用可利用的结构域。这种交错可以通过将结合位点解析成最有用的 3 或 4bp 的结合位点，及通过改变锌指结构域之间的接头长度而实现。为将选择性和高亲和性掺入设计的多肽中，对希望位点具有高特异性的锌指结构域的两侧可以是以较高亲和性但较低特异性结合的其它结构域。所述体内筛选方法可用于测试人工装配的锌指蛋白及其衍生物的体内功能，亲和性及特异性。另外，所述方法可用于最佳化这种装配的蛋白，例如通过产生接头成分，锌指结构域模块，锌指结构域成分等发生变化的文库。

解析靶位点：将 9bp 或更长的靶 DNA 序列解析为 3 或 4bp 的节

段。鉴别识别每个解析的 3 或 4bp 节段的锌指结构域（例如从上述数据库中鉴别）。较长的靶序列，例如 20bp—500bp 的序列，也是合适的靶，因为在其内可鉴别 9bp, 12bp 和 15bp 的亚序列。特别地，可解析为在数据库中充分表现的位点的亚序列可作为初始设计靶。

评分体系可用于评估特别设计的嵌合锌指蛋白识别细胞中靶位点的可能性。所述评分可以是每个锌指成分针对其优选的亚单位的亲和性，其特异性及其在预先设计的蛋白质中的成功的一个函数。

计算机程序：可以使用计算机系统和软件登录上述机读数据库，解析靶位点并输出一或多个嵌合的锌指蛋白设计图。

所述技术可在可编程机器上执行的程序中运行，所述机器如移动或固定计算机及包括处理器，处理器可读的存储介质及一或多个输出设备的相似设备。每个程序均可在高水平程序上或目标定向程序语言中运行，以与机器系统沟通。计算机语言的一些例子包括 C, C++, Java, Fortran 和 Visual Basic 语言。

每个这种程序可以存储于存储介质或设备上，例如光盘只读存储器（CD-ROM），硬盘，磁盘或相似介质或设备上，其是一般或特殊的目的可编程机器可读的，以在当存储介质或设备由计算机读出以进行所述程序时，配置并运行该机器。这个系统也可作为机读存储介质实现，配置一个程序，其中如此配置的存储介质使机器以特异的和预定的方式运行。

所述计算机系统可以与国内或国外网络连接。例如，所述计算机系统可接受来自遥远定位客户系统的请求，例如使用 HTTP, HTTPS 或 XML 协议。该请求可以是已知靶基因的标识符或者代表靶核酸序列的字符串。在前者的情况中，计算机系统可登录一个序列数据库如 GenBank 以检索靶基因调节区域的核酸序列。然后将调节区域的序列或者直接接收的靶核酸序列解析为亚位点，设计例如上述的嵌合锌指蛋白。

该系统可将结果与遥远定位的客户沟通。或者，该系统可控制一个遥控设备以物理检索编码设计的嵌合锌指蛋白的核酸。在这个应用中，构建编码嵌合锌指蛋白的核酸文库并贮存，例如以冷冻的纯化 DNA 或冷冻的携带核酸的细菌菌株形式。所述遥控设备通过登录特异的文库地址回应来自计算机系统的信号。然后可将检索的核酸进行处理，包装并输送至客户。或者，检索的核酸可以导入细胞中并加以分析。然后计算机系统可将分析结果通过网络与客户沟通。

从选择的模块中构建蛋白质：一旦设计了含有多个锌指结构域的嵌合多肽序列，可以合成编码设计的多肽序列的核酸序列。构建合成基因方法是本领域常规方法。这种方法包括从定制合成的寡核苷酸中构建基因，PCR 介导的克隆，及大引物 PCR。下文实施例 66 提供了一种方法，将编码选择的锌指结构域的核酸系列连接以形成编码嵌合多肽的核酸。额外的序列可以与编码设计的多肽序列。所述额外的序列可提供调节功能或者提供编码具有希望功能的氨基酸序列的一个序列。这种额外的序列在此加以描述。

构建嵌合锌指蛋白文库

可以合成编码嵌合锌指蛋白的多个核酸序列，例如以形成文库。编码不同嵌合锌指蛋白的核酸的文库可以通过系列连接而形成，如实施例 67 所述。可构建所述文库由此每个核酸均编码具有至少 3, 4 或 5 个锌指结构域的蛋白质。在一些实施方案中，尤其针对大文库，每个锌指结构域可设计为随机包括一系列锌指结构域中的任一个。可选择一系列锌指结构域以代表具有不同特异性的结构域，例如覆盖 64 个可能的 3bp 的亚位点的 30, 40, 50 或更多个。该系列可包括至少大约 12, 15, 20, 25, 30, 40 或 50 个不同的锌指结构域。一些或所有这些结构域可以是分离自天然存在的蛋白质的结构域。

一个举例文库包括编码嵌合锌指蛋白的核酸，所述锌指蛋白有 3 个指及在每个指上具有 30 个可能的结构域。在其完全表现的形式中，

这个文库包括 27000 个序列（即 30^3 的结果）。这个文库可以通过系列连接而构建，其中编码所有 30 个可能的结构域的核酸集合在每个步骤中加入。最终的文库可作为集合贮存。

或者，可分离各个成员，存储在可寻址位置（例如阵列的）并测序。在高通量测序 4—5 万个构建的文库成员之后，可以单独装配遗漏的嵌合组合物以获得完整覆盖。一旦例如在微滴定平板中形成阵列，每个单独的成员随后均可以回收以进一步分析或者特定应用。特别地，每个单独的成员可通过确定使用实施例 68 所述 p1G 报道基因分析其是否能在体内阻抑转录而确认。如果得以确认，则可以使用核酸微阵列对该文库成员做出分布图，以确定其调节内源基因的能力（见下文“嵌合锌指蛋白调节性质的分布图”）。

小文库，例如具有大约 6—200 个或者 5—2000 个成员，可用于鉴别结合靶位点的最佳嵌合蛋白。这些文库可通过明智选择编码所得嵌合多肽中每个定位的锌指的特定的锌指结构域的核酸组合而设计。例如可选择编码特定位置的核酸加以变化，由此它们编码不同的锌指结构域，其识别特异性适于该位置。

这些小的简明的文库可通过系列连接或通过集合来自预制并列阵的大文库的特异性文库成员而合成。可以使用随后的步骤（例如有性 PCR 及 DNA ShufflingTM（Maxygen 公司，CA）以导入额外的多样性。

筛选嵌合锌指蛋白的文库

所述文库可用特定的筛选程序设计，在这种情况下，亲代载体可工程化为包括必需的调节和功能序列。在一个实施方案中，设计所述文库由此编码嵌合锌指蛋白的核酸两侧为位点特异性重组位点的两侧。重组介导的克隆，例如美国专利 No.5888732 及“Gateway”指导（Life Technologies-Invitrogen，CA，美国）所描述，使每个序列从亲代载体中切下，并插入应用特异性载体中。因此，一旦在亲代载体中

构建了完整的文库，其可以进行不同的筛选和选择程序。

文库成员（来自小或大文库）可以在体内筛选以确定其在细胞中是否能调节感兴趣的靶基因。所述细胞可以是培养的细胞或一个个体体内的细胞。所述靶基因可以是报道基因构建体，其包括与异源报道基因可操纵连接的感兴趣的调节区域，例如实施例 64 所述。或者，所述靶基因可以是一种内源基因。确定由文库成员编码的一或多种蛋白质对内源基因在其正常染色体环境中的调节作用。所述筛选还可包括确定文库成员编码的嵌合蛋白是否改变其它基因的转录。可使用核酸阵列监测大量这种基因的表达，如下文所述。

所述文库还可以使用展示形式筛选，以生物化学性查询文库成员编码的嵌合多肽。例如，文库成员编码的多肽可以展示在噬菌体表面，例如美国专利 5223409 及 Rebar 等（1996），酶学方法 267:129—49 所述。所述文库也可以通过将文库的每个核酸与其编码的多肽共价连接而展示，例如使用 WO 00/32823 所述方法进行。分离具有特定结合性质的各个文库成员可以通过将展示文库与固定在固体支持物上的靶 DNA 位点接触，洗涤所述支持物并回收结合的文库成员而进行。这个方法可适用于鉴别结合其它配体的嵌合多肽，所述其它配体例如是靶 RNA 位点或靶蛋白。

在一个实施方案中，产生每个文库成员编码的嵌合蛋白，并在平面阵列一个地址中分离。产生多肽阵列的方法如以下文献所述：De Wildt 等（2000），自然生物技术 18:989—994；Lueking 等（1999），分析生物化学 270:103—111；Ge, H.（2000），核酸研究 28:e3, I—VII；MacBeath 和 Schreiber（2000），科学 289:1760—1763；Haab 等（2001），基因组生物学 2(2): research0004.1；及 WO99/51773A1。这种阵列可用于鉴别结合特定靶 DNA 位点的文库成员。标记包括靶位点的 DNA 并使之与所述阵列接触。确定在阵列每个地址的标记量以鉴别结合靶位点的文库成员。该分析可包括非特异性 DNA 或者竞

争 DNA 以提高选择的严格性。这个方法可适用于鉴别结合除了 DNA 之外的靶位的嵌合蛋白，例如通过使用标记的靶 RNA 或标记的靶蛋白鉴别。

锌指蛋白的阵列也可以用于对复合的核酸样品作出分布图。将所述样品标记并与所述阵列接触。然后，定量与每个地址的结合以产生该样品的分布图。可将该分布图与参考分布图对比以鉴定所述样品。

嵌合锌指蛋白调节性质的分布图

可对嵌合锌指蛋白进行鉴定以确定其调节细胞的内源基因的能力，所述细胞例如是哺乳动物细胞。首先将编码嵌合锌指蛋白的核酸与阻抑或激活结构域融合，然后导入感兴趣的细胞中。在适当温育并诱导编码核酸表达后，从所述细胞中收获 mRNA 并使用核酸微阵列加以分析。

核酸微阵列可通过各种方法制成，例如照相平版印刷方法（见例如美国专利 No.5510270），机械方法（例如美国专利 No.5384261 所述定向流动方法），基于 pin 的方法（例如美国专利 No.5288514）。合成阵列使得在每个地址有独特的捕捉探针，每个捕捉探针适于检测特定表达的基因的核酸。

mRNA 可以通过常规方法分离，例如包括 DNase 处理以除去基因组 DNA 及与寡 dT 偶联的固相基质杂交（例如分子生物学最新方案所描述，John Wiley&Sons, NY）。洗涤基质以洗脱 mRNA。然后将分离的 mRNA 逆转录及任选例如通过 rtPCR 扩增，例如美国专利 No.4683202 所述。所述核酸可以在扩增或逆转录期间标记，例如通过掺入标记的核苷酸进行。优选的标记例如包括荧光标记，例如红色荧光染料 Cy5（Amersham）或绿色荧光染料 Cy3（Amersham）。或者，所述核酸可以用生物素标记，并在与标记的链亲和素例如链亲和素—藻红蛋白（Molecular Probes）杂交之后检测。

然后将标记的核酸与阵列接触。另外，对照核酸或参考核酸可以

用与样品核酸所用不同的标记加以标记，例如具有不同发射最大值的标记。标记的核酸在杂交条件下与阵列接触。洗涤该阵列，然后成像以检测在阵列的每个地址的荧光。

产生和评价分布图的一般方案包括检测在阵列的每个地址的杂交。在一个地址的杂交程度由一个数值表示并存储在例如一个矢量（vector）中，一个一维矩阵中或者一个一维阵列中。矢量 x 具有阵列每个地址的数值。例如，在特定地址的杂交程度的数值以变量 X_a 存储。所述数值可以被校正，例如针对局部背景水平，样品数量及其它变量校正。核酸也可以从参考样品中制备并与相同或不同的阵列杂交。矢量 y 的构建与矢量 x 相同。可对比样品表达分布图和参考分布图，例如使用数学方程式对比，所述数学方程式是两个矢量的函数。所述对比可作为标量值评估，例如代表两个分布图之间相似性的分值。一或这两个矢量可以通过矩阵转化以在通过阵列检测的不同基因中加入加权值。

表达数据可存储于数据库中，例如相关数据库如 SQL 数据库（例如 Oracle 或 Sybase 数据库环境）。该数据库可具有多个表。例如，原始表达数据可以存储在一个表中，其中每个纵列相当于被分析的基因，例如一个地址或阵列，每个横行相当于一个样品。一个单独的表可存储标识符和样品信息，例如所用阵列的批号，日期及其它质量控制信息。

相似调节的基因可以通过群集表达数据而鉴别，以鉴别共调节的基因。这种群集可以指明嵌合锌指蛋白协同调节的一系列基因。基因可以使用以下方法群集：分级群集（见例如 Sokal 和 Michener (1958), Univ.Kans.Sci.Bull.38: 1409），Bayesian 群集，k-方式群集，及自组织图（见 Tamayo 等（1999），美国科学院院报 96:2907）。

样品表达分布图与参考分布图（例如对照细胞）的相似性也可以如下确定，例如通过将样品表达水平的对数与预测值或参考表达值的

对数进行对比,并通过分布图中所有基因的预测值的加权因子而调节对比。

设计的转录因子的另外特征

肽接头: DNA 结合结构域可以与各种接头连接。本领域熟知接头的用处和设计。特别有用的接头是一种由核酸编码的肽接头。因此,可以构建编码第一个 DNA 结合结构域,肽接头,及第二个 DNA 结合结构域的一个合成基因。可重复这种设计以构建大的合成的多结构域的 DNA 结合蛋白。PCT WO99/45132 及 Kim 和 Pabo ((1998), 美国科学院院报 95:2812-7)描述了适于连接锌指结构域的肽接头的设计。

可利用另外的肽接头,其形成随机卷曲, α -螺旋或 β -折叠三级结构。本领域熟知形成合适的柔性接头的多肽(见例如 Robinson 和 Sauer (1998), 美国科学院院报 95:5929-34)。柔性接头典型包括甘氨酸,因为这个氨基酸没有侧链,在其旋转自由度中是独特的。丝氨酸或苏氨酸可以散布在接头中以提高亲水性。另外,可利用能与 DNA 的磷酸酯主链相互作用的氨基酸,以提高结合亲和性。这种氨基酸的明智使用可平衡亲和性的提高与序列特异性的丧失。如果接头需要刚性延伸,可以使用 α -螺旋接头,如 Pantoliano 等 (1991), 生物化学 30:10117-10125 所述螺旋接头。接头也可以通过计算机建模设计(见例如美国专利 No.4946778)。分子建模软件可商购(例如得自分子模拟公司, San Diego, CA)。接头任选地进行优化例如以降低抗原性和/或提高稳定性,优化使用标准诱变技术及蛋白质工程领域熟知的合适的生物物理学实验及在此所述功能分析进行。

为利用锌指结构域,在锌指之间天然存在的肽可用作接头以将锌指连接在一起。一种典型的这种天然存在的接头是: Thr-Gly-(Glu 或 Gln)-(Lys 或 Arg)-Pro-(Tyr 或 Phe)(SEQ ID NO:78) (Agata 等, 如前)。

二聚体化结构域: 另一种连接 DNA 结构域的方法是使用二聚体

化的结构域，尤其是异源二聚体化结构域（见例如 Pomerantz 等（1998），生物化学 37:965—970）。在这个实施方案中，DNA 结合结构域存在于各独立的多肽链中。例如，第一个多肽编码 DNA 结构域 A，接头和结构域 B，而第二个多肽编码结构域 C，接头和结构域 D。技术人员可从许多充分鉴定的二聚体化结构域中选择一个二聚体化结构域。如果不希望同源二聚体，可以使用异源二聚体化结构域。一个特别合适的二聚体化结构域是卷曲螺旋基序，例如二聚体的平行或反平行的卷曲螺旋。也可获得优先形成异源二聚体的卷曲螺旋序列（Lumb 和 Kim（1995），生物化学 34:8642—8648）。二聚体化结构域的另一端是其中二聚体化通过小分子或信号化触发的结构域。例如，FK506 的二聚体形式可用于二聚体化两个 FK506 结合蛋白（FKBP）结构域。可利用这种二聚体化结构域以提供额外的调节水平。

功能分析及应用

除了生物化学分析之外，通过本发明所述方法例如通过模块装配设计的核酸结合结构域或蛋白质的功能，可以在体内分析或应用。例如，可选择结构域以与靶位点结合，例如与细胞增殖需要的基因的启动子位点结合。通过模块装配，蛋白质可设计为包括（1）分别结合跨越靶启动子位点的亚位点的选择的结构域，（2）DNA 阻抑结构域，例如 WRPW 结构域。

编码设计的蛋白的核酸序列可以克隆入表达载体中，例如 Kang 和 Kim（2000），生物化学杂志 275:8742 所述的可诱导表达载体。所述可诱导表达载体可包括一个诱导型启动子或调节序列。诱导型启动子的非限制性实例包括类固醇激素效应启动子（例如蜕皮激素效应启动子，雌激素效应启动子和糖皮质激素效应启动子），四环素“Tet-On”和“Tet-Off”系统及金属效应启动子。所述构建体可转染入组织培养细胞或胚胎干细胞中以产生转基因生物体作为模型。设计的蛋白质的

效力可以通过诱导蛋白质表达及分析组织培养细胞的增殖或者分析转基因动物模型中发育变化和/或肿瘤生长而确定。另外，被靶向的基因的表达水平可通过常规方法分析以检测 mRNA，例如通过 RT-PCR 或 Northern 印迹进行。更完整的诊断包括从表达或不表达设计的蛋白质的细胞中纯化 mRNA。使用 mRNA 的两个集合探查含有基因大集合的探针的微阵列，所述基因大集合是例如与感兴趣的情况（例如癌症）相关的基因集合或者在生物体基因组中鉴别的基因的集合。这种分析对确定设计的蛋白质的特异性非常有价值。如果蛋白质以高亲和性但低特异性结合，其也许通过影响除了预期的靶基因之外的基因表达而导致多效性和非希望的作用。这种作用通过转录物的整体分析而证实。

另外，设计的蛋白质可以在个体细胞或个体生物体内产生以调节内源基因。如上所述设计的蛋白质的构型使之与内源基因的一个区域结合，并提供转录激活或阻抑功能。如 Kang 和 Kim（如前）所述，编码设计的蛋白质的核酸的表达可以可操纵地与一个诱导型启动子连接。通过调节所述启动子的诱导物的浓度，内源基因的表达可以以浓度依赖方式调节。

分析结合位点优先性

每个结构域的结合位点优先性可以通过生物化学分析而证实，如 EMSA，DNase 足迹试验，表面胞质基因共振或柱结合试验。结合底物可以是涵盖靶位点的合成寡核苷酸。所述分析也可包括非特异性 DNA 作为竞争者或者包括特异性 DNA 序列作为竞争者。特异性竞争 DNA 可包括具有 1，2 或 3 个核苷酸突变的识别位点。因此，可使用生物化学分析不仅测定结构域与给定位点的亲和性，还要测定相对于其它位点其与该位点的亲和性。Rebar 和 Pabo（1994），科学 263:671—673 描述了从 EMSA 中获得锌指结构域的表现 K_d 常数的一种方法。

本发明通过以下实施例得以更详细描述。然而，应注意这些实施

例无限制本发明范围之意。

实施例 1：构建用于杂合转录因子表达的质粒

通过修饰 pPC86 (Chevray 和 Nathans (1991), 美国科学院院报 89:5789—5793) 制备表达锌指转录因子的表达质粒。如 Ausubel 等所述进行 DNA 操作 (分子生物学最新方案 (1998), John Wiley 和 Sons, Inc.)。将编码 Zif268 锌指蛋白的 DNA 片段插入 pPC86 的 SalI 和 EcoRI 识别位点之间以产生 pPCFM-Zif。这个克隆步骤的结果是一种翻译的融合蛋白, 其编码酵母 Gal4 激活结构域后接三个 Zif268 锌指。将 pPCFM-Zif 转化入酵母细胞中导致杂合转录因子表达, 所述杂合转录因子包含酵母 Gal4 激活结构域和 Zif268 锌指。在 pPCFM-Zif 中克隆的编码 Zif268 锌指蛋白的 DNA 序列示于图 9。

质粒 pPCFMS-Zif 用作载体以构建锌指结构域文库 (图 8)。pPCFMS-Zif 通过在 pPCFM-Zif 的第 3 指编码区域之前插入一个寡核苷酸盒而构建, 所述寡核苷酸盒含有一个终止密码子和一个 PstI 识别位点。所述寡核苷酸盒通过退火以下两种合成的寡核苷酸而形成: 5'-TGCCTGCAGCATTTGTGGGAGGAAGTTTG-3' (SEQ ID NO:79); 5'-ATGCTGCAGGCTTAAGGCTTCTCGCCGGTG-3' (SEQ ID NO:80)。插入密码子可防止产生编码 Zif268 的第 3 指的文库质粒。

所述质粒用作产生锌指结构域文库的载体, 如实施例 2 所述。

另外, 如 Hudson 等 ((1997), 基因组研究 7:1169—1173) 所述方法加以略微修改进行编码各个锌指结构域的 DNA 序列的缺口修复克隆。

为克隆各个锌指结构域, 合成两个重叠寡核苷酸。每个寡核苷酸在其 5' 末端包括一个用于进行第二轮 PCR(rePCR) 的长度为 21 个核苷酸的共同尾部, 及一个与编码锌指结构域的核酸退火的特异性序列。正向和反向引物的序列分别是:

5'-ACCCACACTGGCCAGAAACCCN₄₈₋₅₁-3' (SEQ ID NO:108), 及 5'-GATCTGAATTCATTCACCGGTN₄₂₋₄₅-3' (SEQ ID NO:109), 其中 N₄₈₋₅₁ 和 N₄₂₋₄₅ 相当于用于与编码锌指结构域的核酸退火的特定序列。双链 DNA 通过用两种寡核苷酸的等摩尔混合物扩增模板核酸而制备。PCR 条件如下: 在 94℃ 3 分钟第一次循环, 随后在 94℃ 1 分钟, 50℃ 1 分钟及 72℃ 30 秒循环 5 次。

然后将编码每个锌指结构域的双链 DNA 在第二轮 PCR 中用作模板。所述 rePCR 引物有两个区域, 一个区域与酵母载体 pPCFM-Zif 相同, 另一个区域与上述长度为 21 个核苷酸的共同尾部序列相同。正向引物的序列是:

5'-TGTCGAATCTGCATGCGTAACTTCAGTCGTAGTGACCACCTT ACCACCCACATCCGGACCCACACTGGCCAGAAACCC-3' (SEQ ID NO:138), 反向引物的序列是:

5'-GGTGGCGGCCGTTACTTACTTAGAGCTCGACGTCTTACTTACT TAGCGGCCGCACTAGTAGATCTGAATTCATTCACCGGT-3' (SEQ ID NO:139)。反应混合物于 25μl 中含有 2.5 pmol 每种引物, 1.5mM Mg²⁺, 2 单位 Taq 聚合酶和 0.01 单位的 Pfu 聚合酶。反应在 94℃ 进行 3 分钟, 随后在 94℃ 1 分钟, 65℃ 1 分钟及 72℃ 30 秒或者在 94℃ 3 分钟循环 20 次, 然后在 94℃ 30 秒及 72℃ 30 秒循环 25 次。

将 rePCR 产物及已经用 MscI 和 EcoRI 消化的线性化 pPCFM-Zif 载体的混合物转化入酵母 YW1 细胞中进行缺口修复克隆。与酵母载体 pPCFM-Zif 相同的区域使得在酵母细胞中与载体同源重组。由此形成的所有构建体通过 DNA 测序证实。

实施例 2: 构建文库以分析一个锌指结构域

通过从人基因组中克隆锌指结构域制备天然存在的锌指结构域的质粒文库。编码锌指结构域的 DNA 节段从模板人基因组 DNA (购

自 Promega 公司, Madison, WI, 美国) 中扩增, 使用 PCR 和简并的寡核苷酸引物进行。用于克隆人锌指结构域的简并 PCR 引物的 DNA 序列如下: 5'-GCGTCCGGACNCAYACNGGNSARA-3' (SEQ ID NO:81) 及 5'-CGGAATTCANNBRWANGGYTYTC-3' (SEQ ID NO:82), 其中 R 代表 G 和 A; B 代表 G, C 和 T; S 代表 G 和 C; W 代表 A 和 T; Y 代表 C 和 T; N 代表 A, C, G 和 T。

所述简并 PCR 引物与编码氨基酸分布图 His-Thr-Gly-(Glu 或 Gln)-(Lys 或 Arg)-Pro-(Tyr 或 Phe) (SEQ ID NO:83) 的核酸序列退火, 所述氨基酸分布图在许多天然存在的锌指蛋白中的锌指结构域之间连接处发现 (Agata 等 (1998), 基因 213:55—64)。

PCR 反应的缓冲液组合是 50mM KCl, 3mM MgCl₂, 10mM Tris pH8.3。加入 Taq DNA 聚合酶并将反应混合物在 94℃ 30 秒, 在 42℃ 60 秒及在 72℃ 30 秒条件下温育。将此循环重复 35 次, 随后最后在 72℃ 温育 10 分钟。

将 PCR 产物如下克隆入 pPCFMS-Zif 中: 将 PCR 产物电泳并分离大约 120bp 的 DNA 节段。在用 BspEI 和 EcoRI 消化后, 将 120bp 的 DNA 节段连接入 pPCFMS-Zif 中。结果, 由这个质粒文库编码的杂合转录因子的 DNA 结合结构域由 Zif268 的第 1 指和第 2 指及衍生自人基因组的锌指结构域组成。该质粒文库是从共 10⁶ 个大肠杆菌转化体中制备的。这个文库构建方案保留了在锌指结构域之间发现的天然存在的接头序列。

实施例 3: 构建文库以分析一个锌指结构域

通过随机诱变制备突变体锌指结构域文库。Zif268 的第 3 指用作肽构架。沿着 α 螺旋在 -1, 1, 2, 3, 4, 5 和 6 位导入随机突变, 所述位置分别相当于 SEQ ID NO:21 的第 73 位精氨酸, 第 75 位天冬氨酸, 第 76 位谷氨酸, 第 77 位精氨酸, 第 78 位赖氨酸及第 79 位精氨

酸（在 Zif268 的第 3 指内）。

在编码这些氨基酸的每个核酸序列位置中导入一个随机化密码子，5'-(G/A/C)(G/A/C/T)(G/A/C/T)(G/C)-3'。这个随机化密码子编码 16 个氨基酸中的任一个（不包括 4 个氨基酸：色氨酸，酪氨酸，半胱氨酸和苯丙氨酸）。还不包括所有三个可能的终止密码子。所述随机化密码子与一个寡核苷酸盒一起导入，所述寡核苷酸盒从以下两个寡核苷酸中构建：

5'-GGGCCCCGGGGAGAAGCCTTACGCATGTCCAGTCGAATCTTGT
GATAGAAGATTC-3' (SEQ ID NO:84); 及

5'-CTCCCCGCGGTTCGCCGGTGTGGATTCTGATATGSNBSNBAA
GSNBSNBSNBSNBTGAGAATCTTCTATCACAAG-3' (SEQ ID
NO:85)，其中 B 代表 G，T 和 C；S 代表 G 和 C；N 代表 A，G，C
和 T。

在退火这两个寡核苷酸之后，通过与 Klenow 聚合酶反应 30 分钟合成双链 DNA 盒。在用 *Ava*I 和 *Sac*II 消化后，将该双链 DNA 连接入用 *Sgr*AI 和 *Sac*II 消化的 pPCFMS-Zif 中。从大约 109 个大肠杆菌转化体中分离质粒。

实施例 4：构建报道基因质粒

通过修饰 pRS315His (Wang 和 Reed (1993), 自然 364:121-126) 制备包括酵母 HIS3 基因的报道基因质粒。所述报道基因质粒还含有在其天然启动子控制下的 LEU2 标记，以选择携带该质粒的转化体。首先，通过将用 *Sal*I 和 *Bam*HI 消化后的 pRS315His 小片段与用 *Bam*HI 和 *Xho*I 消化后的 pRS315His 大片段连接，从 pRS315His 中除去 *Sal*I 识别位点，产生 pRS315HisΔ*Sal*。接着通过将一个寡核苷酸双链插入在 pRS315HisΔ*Sal* 的 *Bam*HI 和 *Sma*I 位点之间在 HIS3 基因的启动子区域内产生一个新的 *Sal*I 识别位点。用于退火产生插入的双链的两个

寡核苷酸序列是：5'-CTAGACCCGGGAATTCGTCGACG-3' (SEQ ID NO:86)，及 5'-GATCCGTCGACGAATTCCCGGGT-3' (SEQ ID NO:87)。所得质粒称为 pRS315HisMCS。

通过将希望的复合序列插入 pRS315HisMCS 中构建多个报道基因质粒。所述复合序列作为含有复合序列 4 个拷贝的串联阵列插入。靶序列衍生自在 HIV-1 的 LTR 区域中发现的 10bp 的 DNA 序列：

5'-GAC ATC GAG C-3' (SEQ ID NO:1) HIV-1 LTR (-124/-115)

5'-GCA GCT GCT T-3' (SEQ ID NO:2) HIV-1 LTR (-23/-14)

5'-GCT GGG GAC T-3' (SEQ ID NO:3) HIV-1 LTR (-95/-86)

及在人 CCR5 基因的启动子中发现的 10bp 的 DNA 序列：

5'-AGG GTG GAG T-3' (SEQ ID 人 CCR5(-70/-79) NO:4)

5'-GCT GAG ACA T-3' (SEQ ID 人 CCR5(+7/+16) NO:5)。

这些 10bp 的 DNA 序列每个均可以解析为 4bp 的靶位点以鉴别识别该位点每个区域的锌指结构域。使用模块装配方法，这种锌指结构域可以被偶联以产生识别体内位点的 DNA 结合蛋白。

上述序列下划线处描述的是 4bp 的靶序列。将这些 4bp 靶序列的每一个均与 5bp 的募集序列连接，所述募集序列是由 Zif268 的第 1 指和第 2 指识别的序列 5'-GGGCG-3'。所得 9bp 的序列组成复合结合序列。每个复合结合序列有以下格式：5'-XXXXGGGCG-3'，其中 XXXX 是一个 4bp 的靶序列，相邻的 5'-GGGCG-3'是募集序列。

图 7 反映了插入的串联阵列复合结合位点的 DNA 序列，每个 DNA 序列均可操纵地与 pRS315HisMCS 中的报道基因连接。每个串联阵列含有复合结合序列的 4 个拷贝。针对每个结合位点，合成两个寡核苷酸，退火并连接入用 SalI 和 XmaI 位点限制的 pRS315HisMCS 中，产生报道基因质粒。

实施例 5：构建报道基因质粒

如下构建针对每个 3bp 亚位点的一系列报道基因质粒,所述报道基因质粒包括一对报道基因(一个具有 lacZ,另一个具有 HIS3)。报道基因质粒通过将希望的靶序列插入 pRS315HisMCS 和 pLacZi 中而构建。针对每个 3bp 亚位点,合成两个寡核苷酸,退火,并插入 pRS315HisMCS 和 pLacZi 的 SalI 和 XmaI 位点,产生报道基因质粒。所述寡核苷酸的 DNA 序列如下: 5'-CCGGT NNNTGGGCG TAC NNNTGGGCG TCA NNNTGGGCG-3' (SEQ ID NO:88) 和 5'-TCGACGCCCCANN TGA CGCCCCANN GTA CGCCCCANN A-3' (SEQ ID NO:89)。合成共 64 对寡核苷酸并插入两个报道基因质粒中。

实施例 6: 选择具有希望的 DNA 结合特异性的锌指结构域

为选择特异性结合给定靶序列的锌指结构域,将酵母细胞首先用报道基因质粒转化,然后用编码杂合转录因子的杂合质粒文库转化。如 Ausubel 等所述(分子生物学最新方案(1998), John Wiley 和 Sons 公司)进行酵母转化和筛选程序。使用酵母株 yWAM2 (*MAT α* Δ *gal4* Δ *gal80**URA3::GAL1-lacZ lys2801 his3- Δ 200trp1- Δ 63 leu2 ade 2-101CYH2*)。

在一个实例中,将酵母细胞首先用含有可操纵地与报道基因连接的复合结合序列 5'-GAGCGGGCG-3' (下划线处是 4bp 靶序列)的报道基因质粒转化。然后,将通过随机诱变制备的突变体锌指结构域的质粒文库导入转化的酵母细胞中。在没有亮氨酸和色氨酸的培养基中获得大约 10^6 个菌落。由于报道基因质粒和锌指结构域表达质粒分别含有酵母 LEU2 和 TRP1 基因作为标记,因此将酵母细胞在没有亮氨酸和色氨酸的培养基中生长,以选择含有报道基因和锌指结构域表达质粒的细胞。

在一个实施方案中,将衍生自人基因组的锌指结构域文库转化入携带报道基因质粒的细胞中。针对 5 个不同的宿主细胞菌株进行转

化，每个菌株含有与报道基因可操纵地连接的五个不同靶序列之一。在没有亮氨酸和色氨酸的培养基中每次转化获得大约 10^5 个菌落。在温育后，通过向平板中加入 10% 无菌甘油溶液，将菌落刮下置于溶液中并取出该溶液，从而收集转化的细胞。将细胞等份贮存在冷冻的甘油溶液中。将一等份涂布在没有亮氨酸，色氨酸和组氨酸的培养基上。将 3-氨基三唑（AT）加入生长培养基中，终浓度为 0，0.03，0.1 和 0.3mM。AT 是 His3 的一种竞争抑制剂，并代表 HIS3 选择系统的敏感性。AT 抑制 His3 的基础活性。这种基础活性可产生自 HIS3 基因在报道基因质粒上的渗漏表达。除了涂布于培养基上的大约 10^7 个酵母细胞之外，在没有 AT 的选择性培养基中生长几百个菌落。随着 AT 浓度的增加，菌落数量梯度下降。在含有 0.3mM AT 的选择性培养基中生长几十个菌落。从没有 AT 的培养基中及从含有 0.3mM 的 AT 的培养基中随机挑取一些菌落。从这些酵母细胞中分离质粒并转化入大肠杆菌菌株 KC8 (*pyrF leuB600 trpC hisB463*) 中。分离编码锌指转录因子的质粒，确定选择的锌指结构域的 DNA 序列。

每个选择的锌指结构域的氨基酸序列从 DNA 序列中推导。每个锌指结构域以在碱基接触位置的 4 个氨基酸残基命名，所述碱基接触位置沿着 α 螺旋称为 -1，2，3 和 6 位。结果示于表 1。鉴别的锌指结构域由在碱基接触位置发现的 4 个氨基酸命名。序列分析示出在一些情况中重复获得相同的锌指结构域。表 1 中括号中的数字表示相同锌指结构域的获得次数。例如，在 4 个碱基接触位置具有 CSNR 的两个锌指经鉴别结合 GAGC 核酸位点（见第 3 列，GAGC/人基因组）。

表 1

靶序列	GAGC	GAGC	GCTT	GACT	GAGT	ACAT
锌指结构域 文库起源	随机诱变	人基因组	人基因组	人基因组	人基因组	人基因组
在碱基接 触位置的 氨基酸残 基*	KTNR(2) RTTR RPNR HSNR RLKP TRQR TALH RQKA PARV RTFR RNNR DPLH RGNR	RTNR(2) RTNR CSNR(2) SSNR(3) RSTV SSGE	VSTR(9)	HSNK(2) CSNR(7)	RDER(2) SSNR(5)	QSTV(3)

*右侧 6 列中的 4 个字母的标识符是针对每个靶序列分离的锌指结构域的描述符。尽管这些名称表示在碱基接触位置的氨基酸残基，但它们不是多肽序列。

编码选择的人锌指结构域及其翻译的氨基酸序列的全长 DNA 序列示于图 11。下划线处是与简并 PCR 引物互补的 DNA 序列，所述简并 PCR 引物用于扩增编码人基因组中锌指结构域的 DNA 节段。这个序列由于在扩增期间导入的等位基因差异或变化而可与报道的人基因组序列的原始碱基序列不同。

根据本发明所述筛选而鉴别的大多数人锌指结构域是新多肽或者相应于无名开放读框。例如，称为 HSNK（包含于 GenBank 报道的序列中，登记号 AF155100）和 VSTR（包含于 GenBank 报道的序列中，登记号 AF02577）的锌指结构域在还未知其功能的蛋白质中发现。在此所述结果不仅表明这些锌指结构域能发挥序列特异性 DNA

结合结构域的功能，还证明其在嵌合蛋白中优选的结合位点优先性。

另外，本发明揭示了得自人基因组的锌指结构域可用作模块构筑模块，以构建新的 DNA 结合蛋白。本发明的人锌指结构域是作为当与 Zif268 的第 1 指和第 2 指的 C 末端连接时其在体内的功能性的结果获得的。因此，鉴别的锌指结构域可识别人工范围内的特异性序列，并适用作模块构筑模块以设计合成的转录因子。

实施例 7：成对交配

为便于鉴别与每个 3bp 靶位点结合的锌指结构域，用酵母交配消除重复转化酵母细胞的需要，及用单次转化研究 64 个报道基因构建体中每一个的阳性结合子。使用两个酵母株 YW1 (*MAT α* 交配型) 和 YPH499 (*MATa* 交配型)。YW1 通过选择对 5-氟乳清酸 (FOA) 有抗性的克隆以产生 yWAM2 的 *ura3*-衍生物而衍生自 yWAM2。

锌指结构域的质粒文库通过酵母转化导入 YW1 细胞中。用 10% 甘油溶液刮擦平板收集大约 10^6 个独立转化的菌落的细胞。将所述溶液等份冷冻。将 64 个报道基因质粒的每一对 (衍生自 pLacZi 或 pRS315His) 也共转染入酵母株 YPH499 中。收获含有这两种报道基因质粒的转化体并冷冻。

在解冻后，将酵母细胞在基本培养基中生长至中对数期。然后混合两种类型的细胞并使其在 YPD 中交配 5 小时。在含有 X-gal 和 AT (1mM) 但没有色氨酸，亮氨酸，尿嘧啶和组氨酸的基本培养基上选择二倍体细胞。在几天后，分离在选择性平板上生长的蓝色菌落。从蓝色菌落中分离编码锌指结构域的质粒，确定选择的锌指结构域的 DNA 序列。

将分离自蓝色菌落的核酸逐个再转化入 YW1 细胞中。针对每个分离的核酸，将再转化的 YW1 细胞与含有 64 个 LacZ 报道基因质粒的每一个质粒的 YPH499 细胞在 96 孔平板中交配，然后涂布于含有

X-gal 但没有色氨酸和尿嘧啶的基本培养基上。通过蓝色的强度确定锌指结构域与 64 个靶序列的 DNA 结合亲和性和特异性。用 Zif268 锌指结构域进行的对照试验表明锌指结构域和结合位点之间的阳性相互作用产生深蓝色至浅蓝色的菌落，（其蓝色强度与结合亲和性成比例），阴性相互作用产生白色菌落。

实施例 8：鉴别的锌指结构域与相互作用密码的对比

将选择的锌指结构域在关键的碱基接触位置的氨基酸残基与从锌指结构域 DNA 相互作用密码（图 3）中预期的氨基酸残基对比。大多数锌指结构域示出期望的模式，即在关键位置的氨基酸残基与从密码中推测的残基良好匹配。

例如，选自通过随机诱变产生的文库的锌指结构域中的共有氨基酸残基在-1 位是 R（Arg； 7/14）或 K（Lys； 2/14），在第 3 位是 N（Asp； 6/14），及在第 6 位是 R（9/14）（表 1）。这些锌指结构域是用 GAGC 质粒选择的（其中复合结合序列 5'-GAGCGGGCG-3'可操纵地与报道基因连接的报道基因质粒称为 GAGC 质粒。同样，其中序列 5'-XXXXGGGCG-3'可操纵地与报道基因连接的其它报道基因质粒称为 XXXX 质粒）。这些在关键碱基接触位置的氨基酸残基与从密码中预期的那些残基精确匹配。

还已知在第 2 位的氨基酸残基通常仅在碱基识别中发挥微弱作用（Pavletich 和 Pabo（1991）科学 252， 809—817）。然而在一些情况中，第 2 位也许更有影响力。

得自人基因组的锌指结构域中的氨基酸残基也与从密码中预期的残基良好匹配。例如，在用 GAGC 质粒获得的锌指结构域中-1， 3 和 6 位的共有氨基酸残基分别是 R， N 和 R（表 1， 第 3 列）。这些氨基酸是从密码中推测的那些。

在用 GCTT 质粒获得的锌指结构域中第一1， 3 和 6 位的氨基酸

残基分别是 V, T 和 R (表 1, 第 4 列)。所述 T 和 R 残基是从密码子中推测的那些。与 GCTT 位点的 T 碱基 (下划线处) 相互作用的从密码推测的第一 1 位氨基酸残基是 L, T 或 N。用 GCTT 质粒选择的 VSTR 锌指结构域在此位置含有 V (缬氨酸), 这是与 L (亮氨酸) 相似的疏水氨基酸。

总之, 选择的锌指结构域中的氨基酸残基在三个关键位置中的至少两个位置与从密码子推测的那些氨基酸残基匹配。选择的锌指结构域中从密码推测的氨基酸残基在表 1 中是下划线处所示。这些结果强烈提示本文所揭示的体内选择系统如所预期的那样起作用。然而, 因为体内选择和分析系统测定的是细胞中锌指蛋白的实际功能, 它们可鉴别不符合理论期望值 (例如图 3 所示关系) 的具有有用功能和 DNA 结合特异性的锌指。

实施例 9: 再转化及交叉转化

为排除假阳性结果的可能性及研究上述锌指蛋白的序列特异性, 使用分离的质粒进行酵母细胞的再转化和交叉转化。

酵母细胞首先用报道基因质粒和编码锌指结构域的杂合质粒共转化。将酵母转化体接种于没有亮氨酸和色氨酸的基本培养基中并温育 36 小时。将在生长培养基中的大约 1000 个细胞直接点于没有亮氨酸, 色氨酸和组氨酸的固体培养基 (在图 10 中称为 - 组氨酸) 上及点于没有亮氨酸和色氨酸的固体培养基 (在图 10 中称为 + 组氨酸) 上。然后将这些细胞在 30°C 温育 50 小时。结果示于图 10。

预期当杂合转录因子的锌指部分结合复合结合序列时, 菌落可以在没有组氨酸的培养基中生长, 使杂合转录因子激活 HIS3 报道基因的表达。当转录因子的锌指部分不结合复合结合序列时, 菌落在没有组氨酸的培养基中不能生长。

如图 10 所示, 分离的锌指结构域能结合相应的靶序列并示出与

来自 Zif268 的结构域明显不同的序列特异性。Zif268 示出对 GCGT 质粒有比其它五个质粒更高的活性，及对 GAGT 质粒相对较高的活性。具有含有其它结合位点的报道基因并表达 Zif268 蛋白的菌株不形成菌落。

分离自随机突变文库的 KTNR 锌指结构域最初是用 GAGC 报道基因质粒选择的。正如所期望的，只有使用 GAGC 质粒才形成菌落。得自衍生自人基因组的文库的锌指结构域也示出期望的特异性。例如，HSNK，其用 GACT 质粒选择，当再转化入酵母细胞中时示出使只具有 GACT 质粒的细胞生长。VSTR，其用 GCTT 质粒选择，使用 GCTT 质粒示出最高活性。RDER，其用 GAGT 质粒选择，在 4 个碱基接触位置具有与 Zif268 的第 3 指相同的氨基酸残基。正如所期望的，这个锌指结构域示出与第 3 指相似的序列特异性。SSNR，用 GAGC 和 GAGT 质粒选择，使具有 GAGC 质粒但没有 GAGT 质粒的细胞在无组氨酸的培养基上生长。QSTV，用 ACAT 质粒获得，不使具有该分析中测试的任何质粒的细胞生长。然而，这个锌指结构域能与 ACAT 序列在体外紧密结合，如下文所证实。

实施例 10：凝胶移位分析

将含有使用修改的单杂交系统选择的锌指结构域的锌指蛋白在大肠杆菌中表达，纯化，并用于凝胶移位分析中。杂合质粒中编码锌指蛋白的 DNA 节段通过用 SalI 和 NotI 消化而分离，插入 pGEX-4T2 中（Pharmacia Biotech）的 SalI 和 NotI 位点之间。锌指蛋白在大肠杆菌菌株 BL21 中表达为与 GST（谷胱甘肽 S-转移酶）连接的融合蛋白。将所述融合蛋白使用谷胱甘肽亲和性层析（Pharmacia Biotech, Piscataway, NJ）纯化，然后用凝血酶消化，所述凝血酶裂解 GST 部分与锌指蛋白之间的连接位点。纯化的锌指蛋白在 C 末端含有 Zif268 的第 1 指和第 2 指及选择的锌指结构域。

合成以下探针 DNA，退火，使用 T4 多核苷酸激酶用 ^{32}P 标记，并用于凝胶移位分析：

GCGT: 5'-CCGGGTCGCGCGTGGGCGGTACCG-3' (SEQ ID NO:90)

3'-CAGCGCGCACCCGCCATGGCAGCT-5' (SEQ ID NO:91)

GAGC: 5'-CCGGGTCGCGGAGCGGGCGGTACCG-3' (SEQ ID NO:92)

3'-CAGCGCTCGCCCGCCATGGCAGCT-5' (SEQ ID NO:93)

GCTT: 5'-CCGGGTCGTGCTGGGCGGTACCG-3' (SEQ ID NO:94)

3'-CAGCACGAACCCGCCATGGCAGCT-5' (SEQ ID NO:95)

GACT: 5'-CCGGGTCGGGACTGGGCGGTACCG-3' (SEQ ID NO:96)

3'-CAGCCCTGACCCGCCATGGCAGCT-5' (SEQ ID NO:97)

GAGT: 5'-CCGGGTCGGGAGTGGGCGGTACCG-3' (SEQ ID NO:98)

3'-CAGCCCTCACCCGCCATGGCAGCT-5' (SEQ ID NO:99)

ACAT: 5'-CCGGGTCGGACATGGGCGGTACCG-3' (SEQ ID NO:100)

3'-CAGCCTGTACCCGCCATGGCAGCT-5' (SEQ ID NO:101)

将各种数量的锌指蛋白与标记的探针 DNA 在 20mM Tris pH7.7, 120mM NaCl, 5mM MgCl_2 , 20 μM ZnSO_4 , 10%甘油, 0.1% Nonidet P-40, 5mM DTT 和 0.10mg/ml BSA (牛血清白蛋白) 中, 在室温温育 1 小时, 然后将反应混合物进行凝胶电泳。放射性信号通过 PhosphorImagerTM 分析(Molecular Dynamics)加以定量, 解离常数(K_d)如述确定 (Rebar 和 Pabo (1994), 科学 263:671—673)。结果示于表 2。所有常数均在至少两个独立试验中确定, 并标明平均值的标准误差。无组氨酸的基本培养基上酵母转化体的细胞生长 (图 10) 也示于表 2。

表 2

锌指蛋白	探针 DNA	解离常数 (nM)	酵母生长
Zif268	GCTT	2.1±0.3	—
	GCGT	0.024±0.004	+++
	GAGT	0.17±0.04	++
	GAGC	2.3±0.9	—
	GACT	4.9±0.6	—
	ACAT	1.3±0.3	—
KTNR	GCGT	5.5±0.7	—
	GAGC	0.17±0.01	++
	GACT	30±1	—
CSNR	GCGT	2.7±0.3	—
	GAGT	0.46±0.04	+++
	GAGC	1.2±0.1	++
	GACT	0.17±0.01	+++
HSNK	GCGT	42±14	—
	GAGT	3.5±0.1	—
	GACT	0.32±0.08	++
RDER	GCGT	0.027±0.002	+++
	GAGT	0.18±0.01	++
	GACT	28±9	—
SSNR	GCGT	3.8±1.3	—
	GAGC	0.45±0.09	++
	GACT	0.61±0.21	+
VSTR	GCTT	0.53±0.07	++
	GCGT	0.76±0.22	—
	GAGT	1.4±0.2	—
QSTV	GCTT	29±3	—
	GCGT	9.8±3.4	—
	ACAT	2.3±0.4	—

*+++，20—100%生长；++，5—20%生长；+，1—5%生长；—，< 1%生长。

使细胞在无组氨酸的平板上生长的锌指蛋白紧密结合相应的探针 DNA。例如，Zif268 蛋白用作对照，使具有 GCGT 和 GAGT 报道基因质粒的细胞生长，使用相应的探针 DNA 在体外测定的解离常数分别为 0.024nM 和 0.17nM。相反，Zif268 蛋白不使具有其它质粒的细胞生长，使用相应探针 DNA 测定的解离常数高于 1nM。

含有新的锌指结构域的锌指蛋白也示出相似结果。例如，KTNR 蛋白示出与 GAGC 探针 DNA 的强亲和性，解离常数为 0.17nM，但与 GCGT 和 GACT 探针 DNA 没有强亲和性，解离常数分别为 5.5nM 或 30nM。这种蛋白只允许具有 GAGC 质粒的细胞生长。HSNK 蛋白能紧密结合 GACT 探针 DNA ($K_d=0.32\text{nM}$)，但与 GCGT 或 GAGT 探针 DNA 不紧密结合；正如所期望的，HSNK 蛋白只允许具有 GACT 质粒的细胞生长。

QSTV 蛋白，其用 ACAT 报道基因质粒选择，当再转化入酵母细胞中时，不能促进具有任何其它报道基因质粒的细胞生长。凝胶移位分析表明这种蛋白质与其它探针 DNA 相比更紧密结合 ACAT 探针 DNA。即 QSTV 结合 ACAT 探针 DNA 比结合 GCTT 或 GCGT 探针 DNA 分别强 13 倍或 4.3 倍。

通常地，当锌指蛋白例如具有三个锌指结构域的蛋白，结合 DNA 序列的解离常数低于 1nM 时，其允许细胞生长，而当锌指蛋白结合 DNA 序列的解离常数高于 1nM 时，其不允许细胞生长。以解离常数高于 1nM 但低于 5nM 进行结合的锌指蛋白也可以是有用的，例如有 4 个锌指结构域的嵌合锌指蛋白。

实施例 11：TG-ZFD-001 “CSNR1”

TG-ZFD-001 “CSNR1” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCKQCGKAFGCPSNLRRHGRTH (SEQ ID NO:23)。其由下面的人核酸序列编码：

5'-TATAAATGTAAGCAATGTGGGAAAGCTTTTGGATGTCCC
TCAAACCTTCGAAGGCATGGAAGGACTCAC-3' (SEQ ID NO:22)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-001 “CSNR1” 示出对 3bp 靶序列 GAA, GAC 和 GAG 的识别特异性。通过体内筛选结果和 EMSA 确定其结合位点优先性为 GAA>GAC>GAG>GCG。在 EMSA 中, 与 Zif268 的第 1 指和第 2 指及 GST 纯化柄融合的 TG-ZFD-001 “CSNR1”, 对含有 GAC 的位点的表观 K_d 为 0.17nM, 对含有 GAG 的位点的表观 K_d 为 0.46nM, 对含有 GCG 的位点的表观 K_d 为 2.7nM。

TG-ZFD-001 “CSNR1” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GAA, GAC 或 GAG 的 DNA 位点。

实施例 12: TG-ZFD-002 “HSNK”

TG-ZFD-002 “HSNK” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCKEKGKAFNHSSNFNKHHR IH (SEQ ID NO:25)。其由下面的人核酸序列编码:

5'-TATAAGTGTAAGGAGTGTGGGAAAGCCTTCAACCACAG
CTCCA ACTTCAATAAACACCCACAGAATCCAC-3' (SEQ ID NO:24)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-002 “HSNK” 示出对 3bp 靶序列 GAC 的识别特异性。通过体内筛选结果和 EMSA 确定其结合位点优先性为 GAC>GAG>GCG。在 EMSA 中, 与 Zif268 的第 1 指和第 2 指及 GST 纯化柄融合的 TG-ZFD-002 “HSNK”, 对含有 GAC 的位点的表观 K_d 为 0.32nM, 对含有 GAG 的位点的表观 K_d 为 3.5nM, 对含有 GCG 的位点的表观 K_d 为 42nM。

TG-ZFD-002 “HSNK” 可用作模块以构建包含多个锌指结构域

的嵌合 DNA 结合蛋白，例如用于识别含有序列 GAC 的 DNA 位点。

实施例 13: TG-ZFD-003 “SSNR”

TG-ZFD-003 “SSNR” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECKEKGKAFSSGSNFTRHQRIH (SEQ ID NO:27)。其由下面的人核酸序列编码:

5'-TATGAATGTAAGGAATGTGGGAAAGCCTTTAGTAGTGGT
TCAAACCTTCACTCGACATCAGAGAATTCAC-3' (SEQ ID NO:26)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，TG-ZFD-003 “SSNR” 示出对 3bp 靶序列 GAG 的识别特异性。通过体内筛选结果和 EMSA 确定其结合位点优先性为 GAG>GAC>GCG。在 EMSA 中，与 Zif268 的第 1 指和第 2 指及 GST 纯化柄融合的 TG-ZFD-003 “SSNR”，对含有 GAG 的位点的表观 K_d 为 0.45nM，对含有 GAC 的位点的表观 K_d 为 0.61nM，对含有 GCG 的位点的表观 K_d 为 3.8nM。

TG-ZFD-003 “SSNR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GAG 或 GAC 的 DNA 位点。

实施例 14: TG-ZFD-004 “RDER1”

TG-ZFD-004 “RDER1” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YVCDVEGCTWKFARSDELNRHKKRH (SEQ ID NO:29)。其由下面的人核酸序列编码:

5'-TATGTATGCGATGTAGAGGGATGTACGTGGAAATTTGCCC
GCTCAGATGAGCTCAACAGACACAAGAAAAGGCAC-3' (SEQ ID NO:28)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，TG-ZFD-004 “RDER1” 示出对 3bp 靶序列 GCG 的识别特异性。通过体内筛选结果和 EMSA 确定其结合位点优先性为 GCG>GTG, GAG>GAC。在

EMSA 中, 与 Zif268 的第 1 指和第 2 指及 GST 纯化柄融合的 TG-ZFD-004 “RDER1”, 对含有 GCG 的位点的表观 K_d 为 0.027nM, 对含有 GAC 的位点的表观 K_d 为 0.18nM, 对含有 GAC 的位点的表观 K_d 为 28nM。

TG-ZFD-004 “RDER1” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GCG, GTG 或 GAG 的 DNA 位点。

实施例 15: TG-ZFD-005 “QSTV”

TG-ZFD-005 “QSTV” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECNECGKAFAQNSTLRVHQRIH (SEQ ID NO:31)。其由下面的人核酸序列编码:

5'-TATGAGTGTAATGAATGCGGGAAAGCTTTTGCCCAAAT
TCAACTCTCAGAGTACACCAGAGAATTCAC -3' (SEQ ID NO:30)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-005 “QSTV” 示出对 3bp 靶序列 ACA 的识别特异性。通过 EMSA 确定其结合位点优先性为 ACA>GCG>GCT。在 EMSA 中, 与 Zif268 的第 1 指和第 2 指及 GST 纯化柄融合的 TG-ZFD-005 “QSTV”, 对含有 ACA 的位点的表观 K_d 为 2.3nM, 对含有 GCG 的位点的表观 K_d 为 9.8nM, 对含有 GCT 的位点的表观 K_d 为 29nM。

TG-ZFD-005 “QSTV” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 ACA 的 DNA 位点。

实施例 16: TG-ZFD-006 “VSTR”

TG-ZFD-006 “VSTR” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECNYCGKTFSVSSTLIRHQRIH (SEQ ID NO:33)。其由下面的人核酸序列编码:

5'-TATGAGTGTAATTACTGTGGAAAAACCTTTAGTGTGAGC
TCAACCCTTATTAGACATCAGAGAATCCAC-3' (SEQ ID NO:32)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-006 “VSTR” 示出对 3bp 靶序列 GCT 的识别特异性。通过体内筛选结果和 EMSA 确定其结合位点优先性为 GCT>GCG>GAG。在 EMSA 中, 与 Zif268 的第 1 指和第 2 指及 GST 纯化柄融合的 TG-ZFD-006 “VSTR”, 对含有 GCT 的位点的表观 K_d 为 0.53nM, 对含有 GCG 的位点的表观 K_d 为 0.76nM, 对含有 GAG 的位点的表观 K_d 为 1.4nM。

TG-ZFD-006 “VSTR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GCT 或 GCG 的 DNA 位点。

实施例 17: TG-ZFD-007 “CSNR2”

TG-ZFD-007 “CSNR2” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YQCNICGKCFSCNSNLHRHQ RTH (SEQ ID NO:35)。其由下面的人核酸序列编码:

5'-TATCAGTGCAACATTTGCGGAAAATGTTTCTCCTGCAAC
TCCAACCTCCACAGGCACCAGAGAACGCAC-3' (SEQ ID NO:34)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-007 “CSNR2” 示出对 3bp 靶序列 GAA, GAC 和 GAG 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GAA>GAC>GAG。

TG-ZFD-007 “CSNR2” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GAA, GAC 或 GAG 的 DNA 位点。

实施例 18: TG-ZFD-008 “QSHR1”

TG-ZFD-008 “QSHR1” 是通过体内筛选从人基因组序列中鉴别

的。其氨基酸序列为 YACHLCGKAFTQSSHLRRHEKTH (SEQ ID NO:37)。其由下面的人核酸序列编码:

5'-TATGCATGTCATCTATGTGGAAAAGCCTTCACTCAGAGTTCTCACCTTAGAAGACATGAGAAAACCTCAC-3' (SEQ ID NO:36)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-008 “QSHR1” 示出对 3bp 靶序列 GGA, GAA 和 AGA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GGA>GAA>AGA。

TG-ZFD-008 “QSHR1” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GGA, GAA 或 AGA 的 DNA 位点。

实施例 19: TG-ZFD-009 “QSHR2”

TG-ZFD-009 “QSHR2” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCGQCGKFYSQVSHLTRHQKIH (SEQ ID NO:39)。其由下面的人核酸序列编码:

5'-TATAAATGCGGCCAGTGTGGGAAGTTCTACTCGCAGGTC TCCCACCTCACCCGCCACCAGAAAATCCAC-3' (SEQ ID NO:38)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-009 “QSHR2” 示出对 3bp 靶序列 GGA 的识别特异性。

TG-ZFD-009 “QSHR2” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GGA 的 DNA 位点。

实施例 20: TG-ZFD-010 “QSHR3”

TG-ZFD-010 “QSHR3” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YACHLCGKAFTQCSHLRRHEKTH (SEQ ID NO:41)。其由下面的人核酸序列编码:

5'-TATGCATGTCATCTATGTGGAAAAGCCTTCACTCAGTGTT

CTCACCTTAGAAGACATGAGAAAACCTCAC-3' (SEQ ID NO:40)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-010 “QSHR3” 示出对 3bp 靶序列 GGA 和 GAA 的识别特异性。通过体内筛选分析确定其结合位点优先性为 GGA>GAA。

TG-ZFD-010 “QSHR3” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GGA 或 GAA 的 DNA 位点。

实施例 21: TG-ZFD-011 “QSHR4”

TG-ZFD-011 “QSHR4” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YACHLCAKAFIQCSHLRRHEKTH (SEQ ID NO:43)。其由下面的人核酸序列编码:

5'-TATGCATGTCATCTATGTGCAAAAGCCTTCATTCAGTGTTCTCACCTTAGAAGACATGAGAAAACCTCAC-3' (SEQ ID NO:42)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-011 “QSHR4” 示出对 3bp 靶序列 GGA 和 GAA 的识别特异性。通过体内筛选分析确定其结合位点优先性为 GGA>GAA。

TG-ZFD-011 “QSHR4” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GGA 或 GAA 的 DNA 位点。

实施例 22: TG-ZFD-012 “QSHR5”

TG-ZFD-012 “QSHR5” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YVCRECGRGRFRQHS�LVRHKRTH (SEQ ID NO:45)。其由下面的人核酸序列编码:

5'-TATGTTTGCAGGGAATGTGGGCGTGGCTTTCGCCAGCATTCACACCTGGTCAGACACAAGAGGACACAT-3' (SEQ ID NO:44)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-012 “QSHR5” 示出对 3bp 靶序列 GGA, AGA, GAA 和 CGA 的识别特异性。通过体内筛选分析确定其结合位点优先性为 GGA>AGA>GAA>CGA。

TG-ZFD-012 “QSHR5” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GGA, AGA, GAA 或 CGA 的 DNA 位点。

实施例 23: TG-ZFD-013 “QSNR1”

TG-ZFD-013 “QSNR1” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 FECKDCGKAFIQKSNLIRHQ RTH (SEQ ID NO:47)。其由下面的人核酸序列编码:

5'-TTTGAGTGTAAGATTGCGGGAAAGCTTTCATTCAGAAG
TCAAACCTCATCAGACACCAGAGAACTCAC-3' (SEQ ID NO:46)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-013 “QSNR1” 示出对 3bp 靶序列 GAA 的识别特异性。

TG-ZFD-013 “QSNR1” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GAA 的 DNA 位点。

实施例 24: TG-ZFD-014 “QSNR2”

TG-ZFD-014 “QSNR2” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YVCRECRRGFSQKSNLIRHQ RTH (SEQ ID NO:49)。其由下面的人核酸序列编码:

5'-TATGTCTGCAGGGAGTGTAGGCGAGGTTTTAGCCAGAAG
TCAAATCTCATCAGACACCAGAGGACGCAC-3' (SEQ ID NO:48)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-014 “QSNR2” 示出对 3bp 靶序列 GAA 的识别特异性。

TG-ZFD-014 “QSNR2”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白,例如用于识别含有序列 GAA 的 DNA 位点。

实施例 25: TG-ZFD-015 “QSNV1”

TG-ZFD-015 “QSNV1”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECNTRKRTFSQKSNLIVHQRTTH (SEQ ID NO:51)。其由下面的人核酸序列编码:

5'-TATGAATGTAACACATGCAGGAAAACCTTCTCTCAAAG
TCAAATCTCATTGTACATCAGAGAACACAC-3' (SEQ ID NO:50)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-015 “QSNV1”示出对 3bp 靶序列 AAA 和 CAA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 AAA>CAA。

TG-ZFD-015 “QSNV1”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白,例如用于识别含有序列 AAA 或 CAA 的 DNA 位点。

实施例 26: TG-ZFD-016 “QSNV2”

TG-ZFD-016 “QSNV2”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YVCSKCGKAFTQSSNLTVHQKIH (SEQ ID NO:53)。其由下面的人核酸序列编码:

5'-TATGTTTGCTCAAATGTGGGAAAGCCTTCACTCAGAGT
TCAAATCTGACTGTACATCAAAAATCCAC-3' (SEQ ID NO:52)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-016 “QSNV2”示出对 3bp 靶序列 AAA 和 CAA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 AAA>CAA。

TG-ZFD-016 “QSNV2”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白,例如用于识别含有序列 AAA 或 CAA 的 DNA

位点。

实施例 27: TG-ZFD-017 “QSNV3”

TG-ZFD-017 “QSNV3” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCDECGKNFTQSSNLIVHKRIH (SEQ ID NO:55)。其由下面的人核酸序列编码:

5'-TACAAATGTGACGAATGTGGAAAAAAGTTTACCCAGTCC
TCCAACCTTATTGTACATAAGAGAATTCAT-3' (SEQ ID NO:54)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-017 “QSNV3” 示出对 3bp 靶序列 AAA 的识别特异性。

TG-ZFD-017 “QSNV3” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 AAA 的 DNA 位点。

实施例 28: TG-ZFD-018 “QSNV4”

TG-ZFD-018 “QSNV4” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECDVCGKTFTQKSNLGVHQRTTH (SEQ ID NO:57)。其由下面的人核酸序列编码:

5'-TATGAATGTGATGTGTGTGGAAAAACCTTCACGCAAAAG
TCAAACCTTGGTGTACATCAGAGAACTCAT-3' (SEQ ID NO:56)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-018 “QSNV4” 示出对 3bp 靶序列 AAA 的识别特异性。

TG-ZFD-018 “QSNV4” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 AAA 的 DNA 位点。

实施例 29: TG-ZFD-019 “QSSR1”

TG-ZFD-019 “QSSR1” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCPDCGKSFSQSSSLIRHQRTTH (SEQ ID

NO:59)。其由下面的人核酸序列编码：

5'-TATAAGTGCCCTGATTGTGGGAAGAGTTTTAGTCAGAGT
TCCAGCCTCATTCGCCACCAGCGGACACAC-3' (SEQ ID NO:58)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，TG-ZFD-019 “QSSR1” 示出对 3bp 靶序列 GTA 和 GCA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GTA>GCA。

TG-ZFD-019 “QSSR1” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GTA 或 GCA 的 DNA 位点。

实施例 30：TG-ZFD-020 “QSSR2”

TG-ZFD-020 “QSSR2” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECQDCGRAFNQNSSLGRHKRTH (SEQ ID NO:61)。其由下面的人核酸序列编码：

5'-TATGAGTGTCAGGACTGTGGGAGGGCCTTCAACCAGAA
CTCCTCCCTGGGGCGGCACAAGAGGACACAC-3' (SEQ ID NO:60)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，TG-ZFD-020 “QSSR2” 示出对 3bp 靶序列 GTA 的识别特异性。

TG-ZFD-020 “QSSR2” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GTA 的 DNA 位点。

实施例 31：TG-ZFD-021 “QSTR”

TG-ZFD-021 “QSTR” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCEECGKA FNQSSTLTRHKIVH (SEQ ID NO:63)。其由下面的人核酸序列编码：

5'-TACAAATGTGAAGAATGTGGCAAAGCTTTTAACCAGTCC

TCAACCCTTACTAGACATAAGATAGTTCAT-3' (SEQ ID NO:62)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-021 “QSTR” 示出对 3bp 靶序列 GTA 和 GCA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GTA>GCA。

TG-ZFD-021 “QSTR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GTA 或 GCA 的 DNA 位点。

实施例 32: TG-ZFD-022 “RSHR”

TG-ZFD-022 “RSHR” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCMCEGKAFNRRSHLTRHQRIH (SEQ ID NO:65)。其由下面的人核酸序列编码:

5'-TATAAGTGCATGGAGTGTGGGAAGGCTTTTAACCGCAGGTCACACCTCACACGGCACCAGCGGATTCAC-3' (SEQ ID NO:64)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-022 “RSHR” 示出对 3bp 靶序列 GGG 的识别特异性。

TG-ZFD-022 “RSHR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GGG 的 DNA 位点。

实施例 33: TG-ZFD-023 “VSSR”

TG-ZFD-023 “VSSR” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YTCKQCGKAFSVSSSLRRHETTH (SEQ ID NO:67)。其由下面的人核酸序列编码:

5'-TATACATGTAAACAGTGTGGGAAAGCCTTCAGTGTTTCCAGTTCCCTTCGAAGACATGAAACCACTCAC-3' (SEQ ID NO:66)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-023 “VSSR” 示出对 3bp 靶序列 GTT, GTG 和 GTA 的识别特异性。通

过体内筛选结果确定其结合位点优先性为 GTT>GTG>GTA。

TG-ZFD-023 “VSSR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GTT，GTG 或 GTA 的 DNA 位点。

实施例 34: TG-ZFD-024 “QAHR”

TG-ZFD-024 “QAHR” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCKECGQAFRQRAHLIRHHKLH (SEQ ID NO:103)。其由下面的人核酸序列编码：

5'-TATAAGTGTAAGGAATGTGGGCAGGCCTTTAGACAGCGT
GCACATCTTATTCGACATCACAACTTCAC-3' (SEQ ID NO:102)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结果确定 TG-ZFD-024“QAHR”示出对 3bp 靶序列 GGA 的识别特异性。

TG-ZFD-024 “QAHR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GGA 的 DNA 位点。

实施例 35: TG-ZFD-025 “QFNR”

TG-ZFD-025“QFNR”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCHQCGKAFIQSFNLRRHERTH (SEQ ID NO:105)。其由下面的人核酸序列编码：

5'-TATAAGTGTCATCAATGTGGGAAAGCCTTTATTCAATCCT
TTAACCTTCGAAGACATGAGAGAACTCAC-3' (SEQ ID NO:104)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结果确定 TG-ZFD-025“QFNR”示出对 3bp 靶序列 GAG 的识别特异性。

TG-ZFD-025 “QFNR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GAG 的 DNA 位点。

实施例 36: TG-ZFD-026 “QGNR”

TG-ZFD-026 “QGNR” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 FQCNQCGASFTQKGNLLRHIKLH (SEQ ID NO:107)。其由下面的人核酸序列编码:

5'-TTCCAGTGTAATCAGTGTGGGGCATCTTTTACTCAGAAA
GGTAACCTCCTCCGCCACATTAAACTGCAC-3' (SEQ ID NO:106)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, 通过体内筛选结果确定 TG-ZFD-026“QGNR”示出对 3bp 靶序列 GAA 的识别特异性。

TG-ZFD-026 “QGNR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GAA 的 DNA 位点。

实施例 37: TG-ZFD-028 “QSHT”

TG-ZFD-028“QSHT”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCEECGKA FRQSSHLTTHKIIH (SEQ ID NO:111)。其由下面的人核酸序列编码:

5'-TACAAATGTGAAGAATGTGGCAAAGCCTTTAGGCAGTCC
TCACACCTTACTACACATAAGATAATTCAT-3' (SEQ ID NO:110)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-028 “QSHT” 示出对 3bp 靶序列 AGA, CGA, TGA 和 GGA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 (AGA 和 CGA) > TGA > GGA。

TG-ZFD-028 “QSHT” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 AGA, CGA, TGA 和 GGA 的 DNA 位点。

实施例 38: TG-ZFD-029 “QSHV”

TG-ZFD-029“QSHV”是通过体内筛选从人基因组序列中鉴别的。

其氨基酸序列为 YECDHCGKSFSQSSHLNVHKRTH (SEQ ID NO:113)。其由下面的人核酸序列编码:

5'-TATGAGTGTGATCACTGTGGAAAATCCTTTAGCCAGAGC
TCTCATCTGAATGTGCACAAAAGAACTCAC-3' (SEQ ID NO:112)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-029 “QSHV” 示出对 3bp 靶序列 CGA, AGA 和 TGA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 CGA>AGA>TGA。

TG-ZFD-029 “QSHV” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 CGA, AGA 和 TGA 的 DNA 位点。

实施例 39: TG-ZFD-030 “QSNI”

TG-ZFD-030 “QSNI” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YMCSECGRGFSQKSNLIHQRTTH (SEQ ID NO:115)。其由下面的人核酸序列编码:

5'-TACATGTGCAGTGAGTGTGGGCGAGGCTTCAGCCAGAA
GTCAAACCTCATCATACACCAGAGGACACAC-3' (SEQ ID NO:114)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, 通过体内筛选结果确定 TG-ZFD-030 “QSNI” 示出对 3bp 靶序列 AAA 和 CAA 的识别特异性。

TG-ZFD-030 “QSNI” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 AAA 或 CAA 的 DNA 位点。

实施例 40: TG-ZFD-031 “QSNR3”

TG-ZFD-031 “QSNR3” 是通过体内筛选从人基因组序列中鉴别

的。其氨基酸序列为 YECEKCGKAFNQSSNLTRHKKSH (SEQ ID NO:117)。其由下面的人核酸序列编码:

5'-TATGAATGTGAAAAATGTGGCAAAGCTTTTAACCAAGTCC
TCAAATCTTACTAGACATAAGAAAAGTCAT-3' (SEQ ID NO:116)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, 通过体内筛选结果确定 TG-ZFD-031 “QSNR3” 示出对 3bp 靶序列 GAA 的识别特异性。

TG-ZFD-031 “QSNR3” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GAA 的 DNA 位点。

实施例 41: TG-ZFD-032 “QSSR3”

TG-ZFD-032 “QSSR3” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECNECGKFFSQSSSLIRHRRSH (SEQ ID NO:119)。其由下面的人核酸序列编码:

5'-TATGAGTGCAATGAATGTGGGAAGTTTTTTAGCCAGAGC
TCCAGCCTCATTAGACATAGGAGAAGTCAC-3' (SEQ ID NO:118)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-032 “QSSR3” 示出对 3bp 靶序列 GTA 和 GCA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GTA>GCA。

TG-ZFD-032 “QSSR3” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GTA 或 GCA 的 DNA 位点。

实施例 42: TG-ZFD-033 “QTHQ”

TG-ZFD-033 “QTHQ” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECHDCGKSFRQSTHLTQHRRIH (SEQ ID NO:121)。其由下面的人核酸序列编码:

5'-TATGAGTGTCACGATTGCGGAAAGTCCTTTAGGCAGAGC
ACCCACCTCACTCAGCACCGGAGGATCCAC-3' (SEQ ID
NO:120)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-033
“QTHQ”示出对 3bp 靶序列 AGA, TGA 和 CGA 的识别特异性。通
过体内筛选结果确定其结合位点优先性为 AGA>(TGA 和 CGA)。

TG-ZFD-033 “QTHQ”可用作模块以构建包含多个锌指结构域
的嵌合 DNA 结合蛋白, 例如用于识别含有序列 AGA, TGA 和 CGA
的 DNA 位点。

实施例 43: TG-ZFD-034 “QTHR1”

TG-ZFD-034 “QTHR1”是通过体内筛选从人基因组序列中鉴别的。
其氨基酸序列为 YECHDCGKSFRQSTHLTRHRRJH (SEQ ID
NO:123)。其由下面的人核酸序列编码:

5'-TATGAGTGTCACGATTGCGGAAAGTCCTTTAGGCAGAGC
ACCCACCTCACTCGGCACCGGAGGATCCAC-3' (SEQ ID
NO:122)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-034
“QTHR1”示出对 3bp 靶序列 GGA, GAA 和 AGA 的识别特异性。
通过体内筛选结果确定其结合位点优先性为 GGA>(GAA 和 AGA)。

TG-ZFD-034 “QTHR1”可用作模块以构建包含多个锌指结构域
的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GGA, GAA 和 AGA
的 DNA 位点。

实施例 44: TG-ZFD-035 “QTHR2”

TG-ZFD-035 “QTHR2”是通过体内筛选从人基因组序列中鉴别的。
其氨基酸序列为 HKCLECGKCF SQNTHLTRHQRT (SEQ ID

NO:125)。其由下面的人核酸序列编码:

5'-CACAAGTGCCTTGAATGTGGGAAATGCTTCAGTCAGAA
CACCCATCTGACTCGCCACCAACGCACCCAC-3' (SEQ ID
NO:124)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, 通过体内筛选结果确定 TG-ZFD-035 “QTHR2” 示出对 3bp 靶序列 GGA 的识别特异性。

TG-ZFD-035 “QTHR2” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GGA 的 DNA 位点。

实施例 45: TG-ZFD-036 “RDER2”

TG-ZFD-036 “RDER2” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YHCDWDGCGWKFARSDELTRHYRKH (SEQ ID NO:127)。其由下面的人核酸序列编码:

5'-TACCACTGTGACTGGGACGGCTGTGGATGGAAATTCGCC
CGCTCAGATGAACTGACCAGGCACTACCGTAAACAC-3' (SEQ ID
NO:126)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-036 “RDER2” 示出对 3bp 靶序列 GCG 和 GTG 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GCG>GTG。

TG-ZFD-036 “RDER2” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GCG 和 GTG 的 DNA 位点。

实施例 46: TG-ZFD-037 “RDER3”

TG-ZFD-037 “RDER3” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YRCSWEGCEWRFARSDELTRHFRKH (SEQ ID

NO:129)。其由下面的人核酸序列编码：

5'-TACAGATGCTCATGGGAAGGGTGTGGAGTGGCGTTTTGC
AAGAAGTGATGAGTTAACCAGGCACTTCCGAAAGCAC-3' (SEQ
ID NO:128)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结果确定 TG-ZFD-037 “RDER3” 示出对 3bp 靶序列 GCG 和 GTG 的识别特异性。

TG-ZFD-037 “RDER3” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GCG 和 GTG 的 DNA 位点。

实施例 47: TG-ZFD-038 “RDER4”

TG-ZFD-038 “RDER4” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 FSCSWKGCERRFARSDELSRHRRTH (SEQ ID NO:131)。其由下面的人核酸序列编码：

5'-TTCAGCTGTAGCTGGAAAGGTTGTGAAAGGAGGTTTGC
CCGTTCTGATGAACTGTCCAGACACAGGCGAACCCAC-3' (SEQ
ID NO:130)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结果确定 TG-ZFD-038 “RDER4” 示出对 3bp 靶序列 GCG 和 GTG 的识别特异性。

TG-ZFD-038 “RDER4” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GCG 和 GTG 的 DNA 位点。

实施例 48: TG-ZFD-039 “RDER5”

TG-ZFD-039 “RDER5” 是通过体内筛选从人基因组序列中鉴别

的。其氨基酸序列为 FACS WQDCNKKFARSDELARHYRTH (SEQ ID NO:133)。其由下面的人核酸序列编码:

5'-TTCGCCTGCAGCTGGCAGGACTGCAACAAGAAGTTCGC
GCGCTCCGACGAGCTGGCGCGGCACTACCGCACACAC-3' (SEQ
ID NO:132)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, 通过体内筛选结果确定 TG-ZFD-039 “RDER5” 示出对 3bp 靶序列 GCG 的识别特异性。

TG-ZFD-039 “RDER5” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GCG 的 DNA 位点。

实施例 49: TG-ZFD-040 “RDER6”

TG-ZFD-040 “RDER6” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YHCNWDGCGWKFARSDELTRHYRKH (SEQ ID NO:135)。其由下面的人核酸序列编码:

5'-TACCACTGCAACTGGGACGGCTGCGGCTGGAAGTTTGC
GCGCTCAGACGAGCTCACGCGCCACTACCGAAAGCAC-3' (SEQ
ID NO:134)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-040 “RDER6” 示出对 3bp 靶序列 GCG 和 GTG 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GCG>GTG。

TG-ZFD-040 “RDER6” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GCG 和 GTG 的 DNA 位点。

实施例 50: TG-ZFD-041 “RDHR1”

TG-ZFD-041 “RDHR1” 是通过体内筛选从人基因组序列中鉴别

的。其氨基酸序列为 FLCQYCAQRFRGRKDHLTRHMKKSH (SEQ ID NO:137)。其由下面的人核酸序列编码:

5'-TTCCTCTGTCAGTATTGTGCACAGAGATTGTTGGGCGAAAG
GATCACCTGACTCGACATATGAAGAAGAGTCAC-3' (SEQ ID
NO:136)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, 通过体内筛选结果确定 TG-ZFD-041 “RDHR1” 示出对 3bp 靶序列 GAG 和 GGG 的识别特异性。

TG-ZFD-041 “RDHR1” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GAG 和 GGG 的 DNA 位点。

实施例 51: TG-ZFD-043 “RDHT”

TG-ZFD-043 “RDHT” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 FQCKTCQRKFSRSDHLKTHTRTH (SEQ ID NO:141)。其由下面的人核酸序列编码:

5'-TTCCAGTGTA AAACTTGTCAGCGAAAGTTCTCCCGGTCC
GACCACCTGAAGACCCACACCAGGACTCAT-3' (SEQ ID
NO:140)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, 通过体内筛选结果确定 TG-ZFD-043 “RDHT” 示出对 3bp 靶序列 TGG, AGG, CGG 和 GGG 的识别特异性。

TG-ZFD-043 “RDHT” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 TGG, AGG, CGG 和 GGG 的 DNA 位点。

实施例 52: TG-ZFD-044 “RDKI”

TG-ZFD-044“RDKI”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 FACEVCGVRFRNDKCLKIHMRKH (SEQ ID NO:143)。其由下面的人核酸序列编码:

5'-TTTGCCTGCGAGGTCTGCGGTGTTTCGATTCACCAGGAAC
GACAAGCTGAAGATCCACATGCGGAAGCAC-3' (SEQ ID
NO:142)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽,通过体内筛选结果确定 TG-ZFD-044“RDKI”示出对 3bp 靶序列 GGG 的识别特异性。

TG-ZFD-044“RDKI”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白,例如用于识别含有序列 GGG 的 DNA 位点。

实施例 53: TG-ZFD-045 “RDKR”

TG-ZFD-045“RDKR”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YVCDVEGCTWKFARSDKLNHRHKRRH(SEQ ID NO:145)。其由下面的人核酸序列编码:

5'-TATGTATGCGATGTAGAGGGATGTACGTGGAAATTTGCCC
GCTCAGATAAGCTCAACAGACACAAGAAAAGGCAC-3' (SEQ ID
NO:144)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-045“RDKR”示出对 3bp 靶序列 GGG 和 AGG 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GGG>AGG。

TG-ZFD-045“RDKR”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白,例如用于识别含有序列 GGG 和 AGG 的 DNA 位点。

实施例 54: TG-ZFD-046 “RSNR”

TG-ZFD-046“RSNR”是通过体内筛选从人基因组序列中鉴别的。

其氨基酸序列为 YICRKCGRGFSRKS NLIRHQ RTH (SEQ ID NO:147)。其由下面的人核酸序列编码:

5'-TATATTTGCAGAAAGTGTGGACGGGGCTTTAGTCGGAAG
TCCAACCTTATCAGACATCAGAGGACACAC-3'(SEQ ID NO:146)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, TG-ZFD-046 “RSNR” 示出对 3bp 靶序列 GAG 和 GTG 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GAG>GTG。

TG-ZFD-046 “RSNR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GAG 和 GTG 的 DNA 位点。

实施例 55: TG-ZFD-047 “RTNR”

TG-ZFD-047“RTNR”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YLCSECDKCF SRSTNLIRHR RTH (SEQ ID NO:149)。其由下面的人核酸序列编码:

5'-TATCTATGTAGTGAGTGTGACAAATGCTTCAGTAGAAGT
ACAAACCTCATAAGGCATCGAAGAACTCAC-3' (SEQ ID NO:148)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, 通过体内筛选结果确定 TG-ZFD-047“RTNR”示出对 3bp 靶序列 GAG 的识别特异性。

TG-ZFD-047 “RTNR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GAG 的 DNA 位点。

实施例 56: TG-ZFD-048 “HSSR”

TG-ZFD-048“HSSR”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 FKCPVCGKAF RHSSSLVRHQ RTH (SEQ ID NO:173)。其由下面的人核酸序列编码:

5'-TTCAAGTGCCCAAGTGTGCGGCAAGGCCTTCCGGCATAGC
TCCTCGCTGGTGCGGCACCAGCGCACGCAC-3' (SEQ ID
NO:174)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结果确定 TG-ZFD-048 “HSSR” 示出对 3bp 靶序列 GTT 的识别特异性。

TG-ZFD-048 “HSSR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GTT 的 DNA 位点。

实施例 57: TG-ZFD-049 “ISNR”

TG-ZFD-049 “ISNR” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YRCKYCDRSFSISSNLQRHVRNIH (SEQ ID NO:175)。其由下面的人核酸序列编码：

5'-TACAGGTGTAAGTACTGCGACCGCTCCTTCAGCATCTCT
TCGAACCTCCAGCGGCACGTCCGGAACATCCAC-3' (SEQ ID
NO:176)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，TG-ZFD-049 “ISNR” 示出对 3bp 靶序列 GAA, GAT 和 GAC 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GAA>GAT>GAC。

TG-ZFD-049 “ISNR” 可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GAA, GAT 或 GAC 的 DNA 位点。

实施例 58: TG-ZFD-050 “KSNR”

TG-ZFD-050 “KSNR” 是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YGCHLCGKAFSKSSNLRRHEMIH (SEQ ID NO:177)。其由下面的人核酸序列编码：

5'-TATGGATGTCATCTATGTGGGAAAGCCTTCAGTAAAAGT

TCTAACCTTAGACGACATGAGATGATTAC-3' (SEQ ID NO:178)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结果确定 TG-ZFD-050“KSNR”示出对 3bp 靶序列 GAG 的识别特异性。

TG-ZFD-050 “KSNR”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GAG 的 DNA 位点。

实施例 59: TG-ZFD-051 “QSNK”

TG-ZFD-051“QSNK”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YKCEECGKAFTQSSNLTKHKKIH (SEQ ID NO:179)。其由下面的人核酸序列编码：

5'-TACAAGTGTGAAGAATGTGGCAAAGCTTTTACCCAATCC
TCAAACCTTACTAAACATAAGAAAATTCAT-3' (SEQ ID NO:180)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，TG-ZFD-051 “QSNK”示出对 3bp 靶序列 AAA, GAA 和 TAA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GAA>TAA>AAA。

TG-ZFD-051 “QSNK”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GAA, TAA 或 AAA 的 DNA 位点。

实施例 60: TG-ZFD-052 “QSNT”

TG-ZFD-052“QSNT”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECVQCGKGFTQSSNLITHQRVH (SEQ ID NO:181)。其由下面的人核酸序列编码：

5'-TACGAGTGTGTGCAGTGTGGGAAAGGTTTCACCCAGAG
CTCCAACCTCATCACACATCAAAGAGTTCAC-3' (SEQ ID NO:182)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结

果确定 TG-ZFD-052“QSNT”示出对 3bp 靶序列 AAA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 AAA。

TG-ZFD-052 “QSNT”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 AAA 的 DNA 位点。

实施例 61: TG-ZFD-053 “VSNV”

TG-ZFD-053“VSNV”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECDHCGKAFSVSSNLNVHRRJH (SEQ ID NO:183)。其由下面的人核酸序列编码:

5'-TATGAATGCGATCACTGTGGGAAAGCCTTCAGCGTCAGC
TCCAACCTGAACGTGCACAGAAGGATCCAC-3' (SEQ ID
NO:184)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结果确定 TG-ZFD-053 “VSNV”示出对 3bp 靶序列 AAT, CAT 和 TAT 的识别特异性。通过体内筛选结果确定其结合位点优先性为 AAT>CAT>TAT。

TG-ZFD-053 “VSNV”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 AAT, CAT 或 TAT 的 DNA 位点。

实施例 62: TG-ZFD-054 “DSCR”

TG-ZFD-054“DSCR”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YTCSDCGKAFRDKSCLNRHRRTH (SEQ ID NO:185)。其由下面的人核酸序列编码:

5'-TACACATGCAGTGACTGTGGGAAAGCTTTCAGAGATAAA
TCATGTCTCAACAGACATCGGAGAACTCAT-3' (SEQ ID NO:186)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结

果确定 TG-ZFD-054“DSCR”示出对 3bp 靶序列 GCC 的识别特异性。

TG-ZFD-054“DSCR”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GCC 的 DNA 位点。

实施例 63: TG-ZFD-055 “ISNV”

TG-ZFD-055“ISNV”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YECDHCGKAFSIGSNLNVHRRIH (SEQ ID NO:187)。其由下面的人核酸序列编码：

5'-TACGAATGCGATCACTGTGGGAAGGCCTTCAGCATAGGC
TCCAACCTGAATGTGCACAGGCGGATCCAT-3' (SEQ ID NO:188)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结果确定 TG-ZFD-055“ISNV”示出对 3bp 靶序列 AAT 的识别特异性。

TG-ZFD-055“ISNV”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 AAT 的 DNA 位点。

实施例 64: TG-ZFD-056 “WSNR”

TG-ZFD-056“WSNR”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 YRCEECGKAFRWPSNLTRHKRIH (SEQ ID NO:189)。其由下面的人核酸序列编码：

5'-TACAGATGTGAGGAATGTGGCAAAGCCTTTAGGTGGCCC
TCAAACCTTACTAGACATAAGAGAATTTCAC-3' (SEQ ID NO:190)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽，通过体内筛选结果确定 TG-ZFD-056“WSNR”示出对 3bp 靶序列 GGT 和 GGA 的识别特异性。通过体内筛选结果确定其结合位点优先性为 GGT>GGA。

TG-ZFD-056“WSNR”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白，例如用于识别含有序列 GGT 或 GGA 的 DNA 位点。

实施例 65: TG-ZFD-057 “DSAR”

TG-ZFD-057“DSAR”是通过体内筛选从人基因组序列中鉴别的。其氨基酸序列为 FMCTWSYCGKRFTDRSALARHKRTH (SEQ ID NO:191)。其由下面的人核酸序列编码:

5'-TACTCCTGTGGCATTGTGGCAAATCCTTCTCTGACTCCAGTGCCAAAAGGAGACACTGCATTCTACAC-3' (SEQ ID NO:192)。

作为与 Zif268 的第 1 指和第 2 指融合的多肽, 通过体内筛选结果确定 TG-ZFD-057“DSAR”示出对 3bp 靶序列 GTC 的识别特异性。

TG-ZFD-057 “DSAR”可用作模块以构建包含多个锌指结构域的嵌合 DNA 结合蛋白, 例如用于识别含有序列 GTC 的 DNA 位点。

前述锌指结构域及其特异性概括示于下表中。

表3

结合位点	ZFD名称 (SEQ ID NO:)
AAA	QSN1(115), QSNV1(51), QSNV2(53), QSNV3(55), QSNV4(57)
ACA	QSTV(31)
AGA	QSHR1(37), QSHR5(45), QSHT(111), QSHV(113), QTHQ(121), QTHR1(123)
AGG	RDHT(141), RDKR(145)
CAA	QSN1(51), QSNV1(51), QSNV2(53)
CGA	QTHQ(121), QSHR5(45), QSHT(111), QSHV(113)
CGG	RDHT(141)
GAA	CSNR1(23), CSNR2(35), QGNR(107), QSHR1(37), QSHR3(41), QSHR4(43), QSHR5(45), QSNR1(47), QSNR2(49), QSNR3(117), QTHR1(123)
GAC	CSNR1(23), CSNR2(35), HSNK(25), SSNR(27)
GAG	CSNR1(23), CSNR2(35), RDER1(29), RDHR1(137), RSNR(147), RTNR(149), SSNR(27), QFNR(105),
GCA	QSSR1(59), QSSR3(119), QSTR(63)
GCG	RDER1(29), RDER2(127), REDR3(129), RDER4(131), RDER5(133), RDER6(135), VSTR(33)
GCT	VSTR(33)
GGA	QHR1(103), QSHR1(37), QSHR2(39), QSHR3(41), QSHR4(43), QSHR5(45), QSHT(111), QTHR1(123), QTHR2(125)
GGG	RDKI(143), RDHR1(137), RDHT(141), RDKR(143), RSHR(65),
GTA	QSSR1(59), QSSR2(61), QSSR3(119), QSTR(63), VSSR(67)
GTG	RSNR(147), RDER1(29), RDER2(127), RDER3(129), RDER4(131), RDER6(135), VSSR(67)
GTT	VSSR(67)
TGA	QSHT(111), QSHV(113), QTHQ(121)
TGG	RDHT(141)

表4A

结合位点	ZFD名称(SEQ ID NO:)
AAA	QSNK(179), QSNT(181)
AAT	VSNV(183)
CAT	VSNV(183)
GAA	ISNR(175), QSNK(179)
GAC	ISNR(175)
GAG	KSNR(177)
GAT	ISNR(175)
GTT	HSSR(173)
TAA	QSNK(179)
TAT	VSNV(183)

表4B

结合位点	ZFD名称(SEQ ID NO:)
AAT	ISNV (187)
GCC	DSCR (185)
GGT	WSNR (189)
GGA	WSNR (189)
GTC	DSAR (191)

表5

ZFD名称	结合位点	多肽	核酸**
		SEQ ID NO:	SEQ ID NO:
CSNR1	GAA>GAC>GAG	23	22
HSNK	GAC	25	24
SSNR	GAG>GAC	27	26
RDER1	GCG>GTG, GAG	29	28
QSTV	ACA	31	30
VSTR	GCT>GCG	33	32
CSNR2	GAA>GAC>GAG	35	34
QSHR1	GGA>GAA>AGA	37	36
QSHR2	GGA	39	38
QSHR3	GGA>GAA	41	40
QSHR4	GGA>GAA	43	42
QSHR5	GGA>AGA>GAA>CG A	45	44
QSNR1	GAA	47	46
QSNR2	GAA	49	48
QSNV1	AAA>CAA	51	50
QSNV2	AAA>CAA	53	52
QSNV3	AAA	55	54
QSNV4	AAA	57	56
QSSR1	GTA>GCA	59	58
QSSR2	GTA	61	60
QSTR	GTA>GCA	63	62
RSHR	GGG	65	64
VSSR	GTT>GTG>GTA	67	66
QHR	GGA	103	102
QFNR	GAG	105	104
QGNR	GAA	107	106
QSHT	AGA,CGA>TGA>GGA	111	110
QSHV	CGA>AGA>TGA	113	112
QSN1	AAA,CAA	115	114
QSNR3	GAA	117	116
QSSR3	GTA>GCA	119	118
QTHQ	AGA>CGA, TGA	121	120
QTHR1	GGA>GAA, AGA	123	122

QTHR2	GGA	125	124
RDER2	GCG>GTG	127	126
RDER3	GCG>GTG	129	128
RDER4	GCG>GTG	131	130
RDER5	GCG	133	132
RDER6	GCG>GTG	135	134
RDHR1	GAG, GGG	137	136
RDHT	TGG, AGG, CGG, GGG	141	140
RDKI	GGG	143	142
RDKR	GGG>AGG	145	144
RSNR	GAG>GTG	147	146
RTNR	GAG	149	148

**所指出的核酸SEQ ID是指编码锌指结构域的核酸

表6

ZFD名称	结合位点	多肽	核酸**
		SEQ ID NO:	SEQ ID NO:
HSSR	GTT	173	174
ISNR	GAA>GAT>GAC	175	176
KSNR	GAG	177	178
QSNK	GAA>TAA>AAA	179	180
QSNT	AAA	181	182
VSNV	AAT>CAT>TAT	183	184

**所指出的核酸SEQ ID是指编码锌指结构域的核酸

表7

ZFD名称	结合位点	多肽	核酸**
		SEQ ID NO:	SEQ ID NO:
DSCR	GCC	185	186
ISNV	AAT	187	188
WSNR	GGT > GGA	189	190
DSAR	GTC	191	192

**所指出的核酸SEQ ID是指编码锌指结构域的核酸

实施例 66: 构建各个三指蛋白

在哺乳动物细胞中使用载体 P3 表达嵌合锌指蛋白。P3 是通过对 pcDNA3 (Invitrogen, San Diego, CA) 载体加以修饰而构建的。将具有相容突出端的合成寡核苷酸双链连接进用 HindIII 和 XhoI 消化的 pcDNA3 载体中。该双链含有编码血凝素 (HA) 标记及核定位信号的核酸。该双链还包括 BamHI, EcoRI 和 NotI 及 BglII 限制位点及一个终止密码子 (图 11A)。另外, 所得载体的 SV40 起点中 XmaI 位点通过用 XmaI 消化而破坏, 填补消化的限制位点的突出末端, 并将末端再连接。

为构建包括 3 个特定锌指结构域的锌指蛋白, 将编码第一个锌指结构域的核酸与 P3 载体连接。如下所述将编码第二个和第三个锌指结构域的核酸使用 Dynabeads®和 MPC-S 连接在一起。编码第二个和第三个锌指结构域的核酸使用含有 XmaI 位点的正向引物及含有 AgeI 和 NotI 位点的反向引物合成。第二个锌指结构域的正向引物是生物素酰化的。将编码第二个锌指结构域的核酸用 AgeI 消化并与 XmaI 消化的编码第三个锌指结构域的核酸连接。在室温连接 1 小时后, 将连接样品在室温与 Dynabeads®M-280 链亲和素 (DynaI) 结合 15 分钟。用 TE 缓冲液 (10mM Tris HCl, 0.1mM EDTA, pH8.0) 将所述珠洗涤三次。附着的连接样品在 37°C 用 XmaI 和 NotI 消化 3 小时。由 XmaI 和 NotI 消化释放的核酸使用 PCR 纯化试剂盒 (Qiagen) 纯化, 并与包括编码第一个锌指的核酸的 P3 载体连接。将这个最后连接产物转化进大肠杆菌中。鉴别含有在 P3 载体中正确大小的插入体的克隆。编码所得三指 ZFP 的核酸通过 DNA 测序而证实。

实施例 67: 构建三指蛋白文库

图 11B 描述了一种构建多样性三指文库的方法。首先, 将编码每个锌指结构域的核酸克隆入 P3 载体中, 形成“单指”载体。将等

量的每种“单指”载体组合形成集合。将该集合独立地用两个系列的酶消化: AgeI 和 XhoI, 及 XmaI 和 XhoI。在经磷酸酶处理 30 分钟后, 将来自 AgeI 和 XhoI 消化的集合的消化的载体核酸与从 XmaI 和 XhoI 消化的载体中释放的核酸节段连接。这些节段均编码一个单一的锌指结构域。消化的载体核酸与所述核酸节段的连接形成编码两个锌指结构域的载体。在转化入大肠杆菌之后, 所述连接产物产生大约 1.4×10^4 个转化体, 从而形成二指文库。二指文库的插入区域的大小通过对 40 个菌落进行 PCR 分析而查证。在文库中 95% 的是正确大小的插入体。

随后, 将二指文库用 AgeI 和 XhoI 消化。将保留编码两个锌指结构域的核酸序列的消化的载体与上述用 XmaI 和 XhoI 消化而制备的单指片段集合连接。将这一连接的产物转化入大肠杆菌中, 产生大约 2.4×10^5 个独立的转化体。对插入区域进行的查证表明文库成员大多数是正确构建的, 即均编码三个锌指结构域。

实施例 68: 三指蛋白的体内分析

Kim 和 Pabo ((1997), 生物化学杂志 272:29795—29800) 证明当 Zif268 蛋白结合在靶基因的转录起始位点附近时, Zif268 蛋白有效阻抑 VP16 激活的靶基因转录。有假设认为这种结合的锌指蛋白抑制基本转录机制如 RNA 聚合酶 II 复合物与启动子的结合或者 TFIID 与转录起始位点或 TATA 框的结合。

使用相似的体内阻抑分析以确定新的三指蛋白是否在体内起作用。该分析利用萤光素酶报道基因构建体, 其中靶位点位于与 Kim 和 Pabo (如前) 的构建体中 Zif268 位点的位置相应的位置。

萤光素酶报道基因质粒是从 pΔS-modi 中构建的, 这是一种修饰形式的 pGL3-TATA/Inr (Kim 和 Pabo, 如前)。这些报道分子利用萤火虫萤光素酶作为报道蛋白。TATA 框上游中 SacI 位点从 pΔS-modi

中缺失。在转录起始位点之后插入一个新的 SacI 位点。通过 SacI 和 HindIII 消化，用一种含有每个 ZFP 的一个特异的 9bp 结合位点的寡聚物置换位于 pΔS-modi 中转录起始位点下游 12 个碱基后的 14bp。所得报道基因质粒称为 p1G-ZFP ID（ZFP ID 是根据特异性报道基因的结合位点确定的）。pΔS-modi 和 p1G 的序列示于图 12。

如下进行特定的三指蛋白的体内活性分析。将 HEK293 细胞用 4 种质粒转染：14ng 表达特定的三指蛋白的质粒，14ng 上述报道基因质粒，70ng 表达 GAL4-VP16 的质粒，1.4ng 表达 Renilla 萤光素酶的质粒。GAL4-VP16 激活报道基因中最小的合成启动子的转录，由此特定的三指蛋白引起的阻抑作用可以清楚检测，并与其它三指蛋白的作用相对比。表达 Renilla 萤光素酶的质粒提供转染效力对照。

脂转染胺试剂（Gibco-BRL）用于转染程序。将细胞在 96 孔平板的孔中在 30—50%铺满率转染。将细胞温育两天之后收获以进行萤光素酶分析。然后使用 Dual-Luciferase™ 报道基因分析系统（Promega）测定萤光素酶活性。将观测到的萤火虫萤光素酶活性使用 Renilla 的观测水平标准化。阻抑程度或“阻抑倍数”通过将在没有锌指蛋白的情况下标准化的报道基因表达值除以在存在锌指蛋白的情况下标准化的报道基因表达值而计算。

如果在转染分析中锌指蛋白将转录阻抑至少两倍，则将所述锌指蛋白分类为令人满意的高严格度截断值，如果在转染分析中锌指蛋白将转录阻抑 1.5—2 倍，则将所述锌指蛋白分类为低严格度截断值。

实施例 69：ZFPs 与其特异性报道基因的结合分析结果

使用凝胶移位分析将在体内分析中观测的活性与结合亲和性相关联。在通过凝胶移位分析测定的解离常数与在上述转染分析中的转录阻抑水平之间观测到良好关联。表 8 涉及 Zif268 与各种 DNA 位点

的结合亲和性及使用上述转染分析获得的相应体内阻抑数据。

通常地，在转染分析中呈现阻抑大于两倍（即 50%阻抑）的锌指蛋白示出通过凝胶移位分析确定的解离常数低于 1nM。

表 8: Zif268 的体内和体外 DNA 结合活性

靶序列	Zif268 结合的解离常数 (nM)	在表达 Zif268 的人细胞 中的阻抑倍数
GCT TGG GCG	2.1±0.3	1.5±0.1
GCG TGG GCG	0.024±0.004	28±1
GAG TGG GCG	0.17±0.04	5.9±0.1
GAG CGG GCG	2.3±0.9	1.6±0.0
GAC TGG GCG	4.9±0.6	1.2±0.1
ACA TGG GCG	1.3±0.3	1.3±0.1

实施例 70: 三指蛋白的鉴定

构建两种类型的“三指”嵌合锌指蛋白。一种类型包括专门由与天然存在的锌指结构域相同的人锌指结构域组成的嵌合蛋白。另一类型包括包含与天然存在的锌指结构域不同的锌指结构域的嵌合蛋白。后者的锌指结构域通过体外诱变天然存在的锌指结构域随后经噬菌体展示选择而鉴别。这种结构域避免了详细研究天然进化。

锌指结构域的成分示于表 9。共使用 36 个锌指结构域装配一系列测试三指蛋白，其中有 18 个人锌指结构域和 18 个突变的锌指结构域。突变的锌指结构域由以下文献报道：Choo 和 Klug (1994)，美国科学院院报 91:11168—11172；Desjarlais 和 Berg (1994)，美国科学院院报 91:11099—11103；Dreier 等(2001)，生物化学杂志 276:29466—29478；Dreier 等 (2000)，分子生物学杂志 303:489—502；Fairall 等(1993)，自然 366:483—487；Greisman 和 Pabo(1997)，科学 275:657—661；Kim 和 Pabo (1997)，生物化学杂志 272:29795—29800；及 Segal 等 (1999)，美国科学院院报 96:2758—2763。

表9: 用于构建三指ZFP的一些举例性结构域

结构域名称	来源	SEQ ID NO 或参考文献	靶位点
CSNR1	人	23	GAA, GAC, GAG
HSNK	人	25	GAC
ISNR	人	175	GAA, GAT, GAC
QSHR2	人	39	GGA
QSHR3	人	41	GGA, GAA
QSHT	人	111	NGA
QSNR1	人	47	GAA
QSNR3	人	117	GAA
QSNV2	人	53	AAA, CAA
QSSR1	人	59	GTA, GCA
QTHQ	人	121	AGA, CGA, TGA
RDHT	人	141	NGG
RDER1	人	29	GCG, GTG
RSHR	人	65	GGG
RSNR	人	147	GAG, GTG
VSNV	人	183	AAT
VSSR	人	67	GTT, GTG, GTA
VSTR	人	33	GCT
DGAR	突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GTC
DGHR	突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GGC
DGNR	突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GAC
DGNV	突变的/噬菌体展示	J.B.C., Dreier, B. et al. (2001)	AAC
DRDR	突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GCC

GRER 突变的/噬菌体展示	J.B.C., Dreier, B. et al. (2001)	GCC
NDTR 突变的/噬菌体展示	PNAS, Choo & Klug, (1994)	GTT
QAHR 突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GGA
QGDR 突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GCA, GCC, GCT
QGTR 突变的/噬菌体展示	Science, Greisman, H. A. & Pabo, C. (1997)	ACA
QSDR 突变的/噬菌体展示	PNAS, Desjarlais, J. R. & Berg, J. M. (1994)	GCT
QSNR 突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GAA
QSSR 突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GTA
RDKR 突变的/噬菌体展示	J.B.C., Dreier, B. et al. (2001)	GGG
RDNR 突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GAG
RDTN 突变的/噬菌体展示	J.B.C., Dreier, B. et al. (2001)	AAG
TDKR 突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GGG, GGT
TGNR 突变的/噬菌体展示	PNAS, Segal, D. J. et al. (1999)	GAT, GAA

在靶位点的N可以是A、C、G和T中任一个

编码这 36 个结构域的核酸逐个亚克隆入 EcoRI 和 NotI 消化的 P3 载体中, 所得这些质粒用作构建嵌合锌指蛋白的起始材料。

编码嵌合的三指蛋白的核酸通过两种不同的方法制备。

在第一种方法中, 如实施例 68 所述将编码所有锌指结构域的核酸随机混合, 并随机挑取三指构建体进行进一步分析。对每个构建体进行测序以确定其编码的多肽中锌指结构域成分。随后, 针对每个随机混合的三指蛋白合成靶 DNA 序列。靶 DNA 序列基于预期优选的靶位点。将所述靶序列克隆入上述萤光素酶报道基因载体中。这种方法称作“锌指蛋白优先”方法。

在第二个方法中, 将编码嵌合的三指蛋白的核酸基于一给定的靶 DNA 序列而装配。使用计算机程序匹配锌指结构域的识别位点和靶序列。已知基因的启动子序列用作输入靶 DNA 序列。扫描该 DNA 序列以鉴别这样的节段, 即其长度为 9 个核苷酸并且是由嵌合三指蛋白识别的合适靶位点, 提供可利用的锌指结构域集合。一旦鉴别, 构建编码嵌合三指蛋白的核酸。这种方法称作“靶位点优先”方法。

根据特殊考虑, 分析在碱基接触残基第二位包括天冬氨酸残基的锌指结构域。这种锌指结构域包括 RDER1, RDHT, RDNR, RDKR, RDTN, TDKR 及 NDTR。与 DNA 结合的 Zif268 的 X 线共结晶结构示出在第二位的天冬氨酸可以与锌指识别的 3bp 的亚位点外侧的一个碱基形成一个氢键。结果, 在第二位含有天冬氨酸残基的 RDER 锌指优选 4bp 位点: 5'-GCG(G/T)-3'。计算机算法说明了这种额外的特异性。相似地, 在本文所述的其它分析中不包括如下的随机装配的三指蛋白, 所述三指蛋白包括在第二位具有天冬氨酸的锌指并破坏了 4bp 位点规则。

从“锌指蛋白优先”和“靶位点优先”方法中共构建了 153 个三指嵌合蛋白。使用实施例 68 所述瞬时共转染方法测试这些蛋白。结果示于表 10。

表10: 锌指蛋白及其DNA结合活性

	结构域名称			靶复合序列	阻抑 倍数
	1	2	3		
1	RSNR	RDHT	RSNR	5'-GAG AGG GAG C-3' (SEQ ID NO:199)	8.1
2	CSNR1	RSHR	RDHT	5'-TGG GGG GAC A-3' (SEQ ID NO:200)	4.5
3	RDHT	QSNV2	QSSR1	5'-GCA CAA TGG-3'	4.0
4	RSHR	RDER1	RDER1	5'-GCG GCG GGG C-3' (SEQ ID NO:201)	3.8
5	QSHR3	QAGR*	QSSR*	5'-GTA GGA GGA T-3' (SEQ ID NO:202)	3.7
6	RSHR	RDER1	RDHT	5'-AGG GCG GGG C-3' (SEQ ID NO:203)	3.6
7	RDHT	QSNV2	RSHR	5'-GGG AAA CGG G-3' (SEQ ID NO:204)	3.6
8	RSNR	QSHR2	QSSR1	5'-GTA GGA GAG T-3' (SEQ ID NO:205)	3.3
9	QSDR*	RDHT	QSHR3	5'-GGA AGG GCT T-3' (SEQ ID NO:206)	3.3
10	RDHT	RSHR	QSHR2	5- GGA GGG TGG -3'	3.1
11	QSHT	RSHR	RDHT	5'-TGG GGG TGA-3'	3.0
12	QSSR1	QSNV2	RSNR	5'-GAG CAA GTA G-3' (SEQ ID NO:207)	2.9
13	QSHR2	RDER1	RSNR	5'-GAG GTG GGA G-3' (SEQ ID NO:208)	2.8
14	DGHR*	QSNR1	RDHT	5'-GTA GGC GCC-3'	2.6
15	VSNV	CSNR1	RSNR	5'-GAG GAC AAT G -3' (SEQ ID NO:209)	2.5
16	QSHR2	RDER1	RSHR	5'-GGG GCG GGA T-3' (SEQ ID NO:210)	2.3
17	VSSR	QGDR*	RDHT	5'-GGG GCA GTT-3'	2.3
18	RDER1	QSSR1	QSHT	5'-CGA GCA GCG-3'	2.2
19	RDER1	VSSR	QSHT	5'-CGA GTT GCG-3'	2.1

20	QSNR3	QSHR2	RSNR	5'-GAG GGA GAA G-3' (SEQ ID NO:211)	2.1
21	RDHT	QSHR	RSNR	5'-GGA GAG AGA-3'	2.1
22	RSNR	VSSR	RSHR	5'-GGG GTT GAG-3'	2.1
23	RDHT	RSNR	QSNR3	5'-GAA GAG AGG T-3' (SEQ ID NO:212)	2.1
24	CSNR1	QSHR	RSNR	5'-GAG TGA GAC C-3' (SEQ ID NO:213)	2.1
25	QSNR3	RDER1	RSNR	5'-GAG GCG GAA A3' (SEQ ID NO:214)	2.1
26	QSNR3	QSNV2	RSHR	5'-GGG AAA GAA C-3' (SEQ ID NO:215)	2.0
27	RSHR	RDER1	RSHR	5'-GGG GTG GGG-3'	2.0
28	VSSR	QSNR3	RDER1	5'-GCG GAA GTT C-3' (SEQ ID NO:216)	2.0
29	QSNR3	RDHT	RSNR	5'-GAG TGG GAA A-3' (SEQ ID NO:217)	2.0
30	RSHR	RSHR	QSHR2	5'-GGA GGG GGG C-3' (SEQ ID NO:218)	2.0
31	ISNR	RSNR	RDHT	5'-TGG GAG GAT C-3' (SEQ ID NO:219)	2.0
32	QSNV2	RSNR	RDHT	5'-GGG GAG AAA-3'	1.9
33	QSNV2	RSHR	RDER1	5'-GTG GGG AAA A-3' (SEQ ID NO:220)	1.9
34	VSSR	HSNK	RSNR	5'-GAG GAC GTG -3'	1.9
35	VSSR	QSHR	RSHR	5'-GGG TGA GTG-3'	1.9
36	RSNR	VSSR	RSNR	5'-GAG GTT GAG G-3' (SEQ ID NO:221)	1.8
37	QAHR*	QSNR	QTHQ	5'-AGA GAA GGA G-3' (SEQ ID NO:222)	1.8
38	RSNR	ISNR	QSHR	5'-TGA GAT GAG C-3' (SEQ ID NO:223)	1.8
39	VSTR	RSNR	QSHR2	5'-GGA GAG GCT C-3' (SEQ ID NO:224)	1.8
40	ISNR	VSTR	RDHT	5'-AGG GCT GAT T-3' (SEQ ID NO:225)	1.8

41	QSNR3	RSNR	RSHR	5'-GGG GAG GAA A-3' (SEQ ID NO:226)	1.7
42	RDHT	QSHR2	QSHT	5'-AGA GGA AGG T-3' (SEQ ID NO:227)	1.7
43	QSSR1	QSNR3	QSHR2	5'-GGA GAA GTA G-3' (SEQ ID NO:228)	1.7
44	RDHT	DGHR*	TDKR*	5'-GGT GGC AGG T-3' (SEQ ID NO:229)	1.7
45	HSNK	QSNV2	VSSR	5'-GTT CAA GAC-3'	1.7
46	HSNK	QSHT	RSNR	5'-GAG AGA GAC-3'	1.7
47	RSHR	QSHR2	VSTR	5'-GCT GGA GGG G-3' (SEQ ID NO:230)	1.7
48	RDHT	RSHR	RDER1	5'-GCG GGG AGG G-3' (SEQ ID NO:231)	1.6
49	RSHR	QSNV2	QSHT	5'-AGA AAA GGG-3'	1.6
50	RSHR	RDER1	QSNV2	5'-AAA GTG GGG A-3' (SEQ ID NO:232)	1.6
51	VSNV	QSNV2	QSHT	5'-AGA AAA AAT A-3' (SEQ ID NO:233)	1.6
52	QSHR2	QSHT	RSNR	5'-GAG TGA GGA-3'	1.6
53	QSHR2	RDHT	CSNR1	5'-GAC AGG GGA G-3' (SEQ ID NO:234)	1.6
54	QSHT	VSSR	RSHR	5'-TGA GTT GGG A-3' (SEQ ID NO:235)	1.6
55	QSNV2	QSHR2	QSNR3	5'-GAA GGA AAA T-3' (SEQ ID NO:236)	1.6
56	RSNR	VSTR	RSHR	5'-GGG GCT GAG G-3' (SEQ ID NO:237)	1.5
57	QSHR2	CSNR1	QSHT	5'-TGA GAC GGA G-3' (SEQ ID NO:238)	1.5
58	VSNV	QSHR2	VSTR	5'-GCT GGA AAT T-3' (SEQ ID NO:239)	1.5
59	DGAR*	DGNR*	RDKR*	5'-GGG GAC GTC-3'	1.5
60	QSNR3	QSSR1	QSNV2	5'-CAA GTA GAA G-3' (SEQ ID NO:240)	1.5
61	QSNR3	RDER1	RSNR	5'-GAG GCG GAA A-3'	1.5

			(SEQ ID NO:241)	
62	RDER1	QSSR1	VSTR 5'-GCT GCA GCG T-3' (SEQ ID NO:242)	1.5
63	QSHR3	DGHR*	RDHT 5'-GGG GGC GGA-3'	1.5
64	VSSR	RSHR	ISNR 5'-GAT GGG GTT T-3' (SEQ ID NO:243)	1.5
65	RSNR	RDER1	RSHR 5'-GGG GCG GAG-3'	1.5
66	RSNR	RDER1	QSNR3 5'-GAA GCG GAG G-3' (SEQ ID NO:244)	1.4
67	QSHR3	QSNV2	DGHR* 5'-GGC AAA GGA-3'	1.4
68	QSNV2	QSNR3	QSSR1 5'-GTA GAA AAA-3'	1.4
69	QSNR3	RDER1	RDER1 5'-GTG GCG GAA G-3' (SEQ ID NO:245)	1.4
70	QSNV2	QSDR*	QSSR1 5'-GTA GCT AAA-3'	1.4
71	QSNV2	QSHR2	QSNV2 5'-AAA GGA AAA G-3' (SEQ ID NO:246)	1.4
72	QSNV2	VSSR	RDHT 5'-CGG GTT AAA A-3' (SEQ ID NO:247)	1.4
73	QSHR3	QTHQ	QSDR* 5'-GCT AGA GGA-3'	1.4
74	RSHR	VSTR	QSSR1 5'-GTA GCT GGG A-3' (SEQ ID NO:248)	1.4
75	RDER1	QSNV2	QSHR2 5'-GGA CAA GCG G-3' (SEQ ID NO:249)	1.3
76	QSNV2	QSHT	QSNV2 5'-AAA AGA AAA A-3' (SEQ ID NO:250)	1.3
77	QSHR2	QSHR2	QSSR1 5'-GTA GGA GGA T-3' (SEQ ID NO:202)	1.3
78	QSNR3	QSSR1	QSHT 5'-AGA GTA GAA T-3' (SEQ ID NO:252)	1.3
79	QSNV2	QSSR1	QSNV2 5'-AAA GTA AAA A-3' (SEQ ID NO:253)	1.3
80	QSHR2	RSNR	RDHT 5'-AGG GAG GGA G-3' (SEQ ID NO:254)	1.3
81	RSNR	VSNV	QSNV2 5'-AAA AAT GAG C-3' (SEQ ID NO:255)	1.3
82	QSHT	QSNR3	RDHT 5'-CGG GAA AGA A-3'	1.3

			(SEQ ID NO:256)	
83	QTHQ	QGDR*	QSSR* 5'-GTA GCA AGA C-3' (SEQ ID NO:257)	1.3
84	QSNV2	QSSR1	VSNV 5'-AAT GTA AAA A-3' (SEQ ID NO:258)	1.3
85	RDKR*	QAHR*	RDHT 5'-CGG GGA GGG G-3' (SEQ ID NO:259)	1.3
86	QSNR*	QSNR*	QAHR* 5'-TTC TTC TCC-3'	1.3
87	DGNR*	RSNR	QSSR1 5'-GTA GAG GAC-3'	1.2
88	CSNR1	QSHT	QSNV2 5'-CAA AGA GAC T-3' (SEQ ID NO:260)	1.2
89	RDER1	ISNR	QSNR3 5'-GAA GAT GCG T-3' (SEQ ID NO:261)	1.2
90	RDHT	QSSR1	QSHT 5'-CGA GCA TGG G-3' (SEQ ID NO:262)	1.2
91	RDHT	QGTR*	QGTR* 5'-ACA ACA GGG G-3' (SEQ ID NO:263)	1.20
92	RSHR	RSHR	VSSR 5'-GTT GGG GGG C-3' (SEQ ID NO:264)	1.2
93	RDER1	RSNR	RDHT 5'-AGG GAG GTG T-3' (SEQ ID NO:265)	1.2
94	RSHR	CSNR1	QSHT 5'-TGA GAC GGG G-3' (SEQ ID NO:266)	1.2
95	QSHR2	VSSR	QSNR3 5'-GAA GTT GGA A-3' (SEQ ID NO:267)	1.2
96	QSNR3	QSNV2	QSHT 5'-AGA AAA GAA A-3' (SEQ ID NO:268)	1.2
97	QSNV2	QSHT	CSNR1 5'-GAC TGA CAA T-3' (SEQ ID NO:269)	1.2
98	TGNR*	RDNR*	QSDR* 5'-GCT GAG GAT G-3' (SEQ ID NO:270)	1.2
99	QSNV2	RSNR	RSHR 5'-GGG GAG AAA T-3' (SEQ ID NO:271)	1.2
100	QSNR3	QSHT	QSHT 5'-TGA TGA GAA A-3' (SEQ ID NO:272)	1.2
101	HSNK	QSHR2	QSSR1 5'-GCA GGA GAC T-3'	1.2

			(SEQ ID NO:273)	
102	TGNR*	QAHR*	RDHT 5'-TGG GGA GAT T-3' (SEQ ID NO:274)	1.1
103	RDHT	QSNR3	RDER1 5'-GCG GAA TGG A-3' (SEQ ID NO:275)	1.1
104	QSHR3	RDHT	DGAR* 5'-GTC TGG GGA C-3' (SEQ ID NO:276)	1.1
105	RDER1	RSHR	RSNR 5'-GAG GGG GCG T-3' (SEQ ID NO:277)	1.1
106	VSTR	VSTR	QSSR1 5'-GAC GCT GCT T-3' (SEQ ID NO:278)	1.1
107	TGNR*	QAHR*	QAHR* 5'-ATC TCC TCC-3'	1.1
108	DGHR*	QGDR*	RDKR* 5'-GGG GCA GGC G-3' (SEQ ID NO:279)	1.1
109	RSHR	RSHR	TGNR* 5'-GAT GGG GGG-3'	1.1
110	QSNV2	QSNV2	QSNV2 5'-AAA AAA AAA G-3' (SEQ ID NO:280)	1.1
111	RSNR	QSHT	QSHR2 5'-GGA AGA GAG G-3' (SEQ ID NO:281)	1.1
112	QSNV2	RSHR	QSNV2 5'-CAA GGG AAA A-3' (SEQ ID NO:282)	1.1
113	QGDR*	TGNR*	TDKR* 5'-GGT .GAT GCA C-3' (SEQ ID NO:283)	1.1
114	RDER1	DGAR*	RDTN* 5'-AAG GTC GCG G-3' (SEQ ID NO:284)	1.0
115	QAHR*	QSDR*	RDKR* 5'-GGG GCT GGA G-3' (SEQ ID NO:285)	1.0
116	VSSR	TDKR*	RDHT 5'-GGG GGG GTT-3'	1.0
117	QSSR*	TDKR*	RDKR* 5'-GGG GGT GTA C-3' (SEQ ID NO:286)	1.0
118	QSDR*	TGNR*	TDKR* 5'-GGT GAT GCT C-3' (SEQ ID NO:287)	1.0
119	CSNR1	QSHT	VSSR 5'-GTT TGA GAC A-3' (SEQ ID NO:288)	1.0
120	RSNR	QSHR3	DGHR* 5'-GGC GGA GAG-3'	1.0
121	VSNV	QSNV2	VSTR 5'-GCT AAA AAT C-3'	1.0

			(SEQ ID NO:289)	
122	QSDR*	QAHR*	QTHQ 5'-AGA GGA GCT T-3' (SEQ ID NO:290)	1.0
123	RSHR	ISNR	QSHT 5'-TGA GAT GGG G-3' (SEQ ID NO:291)	1.0
124	QSNR*	QAHR*	RDKR* 5'-TTC TCC CCC-3'	0.9
125	DGHR*	RDHT	QSDR* 5'-GCT TGG GGC T-3' (SEQ ID NO:292)	0.9
126	RDER1	QGDR	VSSR 5'-GTT GGG GCG G-3' (SEQ ID NO:293)	0.9
127	QSDR*	QSDR*	QAHR* 5'-GGA GCT GCT T-3' (SEQ ID NO:294)	0.9
128	DGHR*	RSNR	DGNV* 5'-AAC GAG GGC-3'	0.9
129	QSHR3	QGDR*	TGNR* 5'-GAT GCA GGA C-3' (SEQ ID NO:295)	0.9
130	QSNR1	DRDR*	TGNR* 5'-GAT GCC GAA-3'	0.9
131	DGAR*	RDHT	DGHR* 5'-GGC CGG GTC G-3' (SEQ ID NO:296)	0.9
132	TDKR*	TDKR*	TGNR* 5'-GAT GGT GGT T-3' (SEQ ID NO:297)	0.9
133	NDTR*	QSNR*	QAHR* 5'-GGA GAA GTT-3'	0.9
134	TGNR*	QGTR*	QSDR* 5'-GCT ACA GAT-3'	0.8
135	RDER1	TDKR*	DRDR* 5'-GCC GGG GCG G-3' (SEQ ID NO:298)	0.8
136	QSNR1	VSSR	DGNV* 5'-AAC GTT GAA-3'	0.8
137	DGAR*	QSSR*	GRER* 5'-GCC GTA GTC-3'	0.8
138	CSNR1	RSHR	VSTR 5'-GCT GGG GAC T-3' (SEQ ID NO:299)	0.8
139	QSSR*	QSDR*	QSSR* 5'-GTA GCT GTA A-3' (SEQ ID NO:300)	0.8
140	QSNV2	RSNR	QSSR1 5'-GTA GAG AAA-3'	0.8
141	TDKR*	DGHR*	RDKR* 5'-GGG GGC GGT T-3' (SEQ ID NO:301)	0.8
142	TGNR*	QSDR*	TDKR* 5'-GGT GCT GAT T-3' (SEQ ID NO:302)	0.8
143	DRDR*	DGHR*	QSSR1 5'-GTA GGC GCC-3'	0.8

144	QAHR*	QSSR*	QGDR*	5'-GCA GTA GGA G-3' (SEQ ID NO:303)	0.8
145	QSHT	RSNR	QSDR*	5'-GCT GAG AGA-3'	0.8
146	DGNV*	QGDR*	DGNV*	5'-AAC GCA AAC-3'	0.7
147	QGDR*	TDKR*	DGAR*	5'-GTC GGG GCA-3'	0.7
148	TDKR*	QSNR1	DGNR*	5'-GAC GAA GGG G-3' (SEQ ID NO:304)	0.7
149	QSNR*	QGDR*	NDTR*	5'-GTT GCT GAA-3'	0.7
150	QSNR*	RDKR*	DGHR*	5'-GGC GGG GAA-3'	0.7
151	DGNV*	QGDR*	DGNR*	5'-GAC GCA AAC-3'	0.7
152	QSDR*	DGNR*	DGNR*	5'-GAC GAC GCT T-3' (SEQ ID NO:305)	0.7
153	CSNR1	TGNR*	DGNR*	5'-GAC GAT GAA-3'	0.6

后面有“*”的结构域是通过诱变获得的锌指结构域。未如此标示的结构域是本发明所述人锌指结构域。靶复合序列通过并列每个结构域的各个靶位点而设计。

与高和低严格度标准相关的结果的分布示于表 11 和表 12。如表 11 所示，153 个嵌合锌指蛋白中有 31 个表明高于 2 倍阻抑，高严格标准（RF 2；RF=阻抑倍数）。表 12 表明在完全从天然存在的人锌指结构域中构建的蛋白质中 28.1%（27/96）超过高严格标准，59.4% 超过低严格标准（RF 1.5）。在从两个天然存在的锌指结构域和一个突变结构域中构建的蛋白质中，33.3% 超过高严格标准，只有 20% 超过低严格标准。

相反，在从一个人结构域和两个突变结构域中构建的 17 种蛋白质中，只有一种蛋白质（5.9%）超过高严格标准，只有两个（11.8%）超过低严格标准。只由突变结构域组成的锌指蛋白在阻抑分析中不符合高严格标准。这种蛋白质只有一个（4%）符合低严格标准。这些结果表明天然存在的人锌指结构域在构建新的 DNA 结合蛋白中与突变的结构域相比是更好的构筑模块。

表 11

测试 ZFP 的数量 (A)	活性 ZFP 的数量 (B)		B/A (%)	
	RF>1.5	RF>2.0	RF>1.5	RF>2.0
153	65	31	42.5	20.3

表 12

ZFP 的组成		测试的 ZFP 数量 (A)	活性 ZFP 数量 (B)		B/A (%)	
人结构 域 D	突变结 构域 D		RF>1.5	RF>2.0	RF>1.5	RF>2.0
3	0	96	57	27	59.4	28.1
2	1	15	5	3	33.3	20
1	2	17	2	1	11.8	5.9
0	3	25	1	0	4.0	0

本文描述了本发明的许多实施方案。然而应理解在不偏离本发明精神和范围的前提下可对本发明加以各种修改。因此，其它实施方案也在所附权利要求的范围内。

序列表

<110> 图尔金株式会社

<120> 锌指结构域文库

<130> A20952/TGI

<150> US60/374,355

<151> 2002-04-22

<150> US60/313,402

<151> 2001-08-17

<160> 305

<170> FastSEQ for Windows Version 4.0

<210> 1

<211> 10

<212> DNA

<213> HIV-1

<400> 1

gacatcgagc

10

<210> 2

<211> 10

<212> DNA

<213> HIV-1

<400> 2

gcagctgctt

10

<210> 3

<211> 10

<212> DNA

<213> HIV-1

<400> 3

gctggggact

10

<210> 4

<211> 10

<212> DNA

<213> Homo sapiens

<400> 4

agggtggagt

10

<210> 5

<211> 10

<212> DNA

<213> Homo sapiens

<400> 5

gctgagacat

10

<210> 6

<211> 47

<212> DNA

<213> 人工序列

<220>

<223> 最佳结合位点

<400> 6

ccggcgtgggcggcgtgcgtgggcgtgcgtgggcggactgc gtgggcg

47

<210> 7

<211> 47

<212> DNA

<213> 人工序列

<220>

<223> 最佳结合位点

<400> 7

tcgacgccacgcagtcgccacgcacgccacgcagccgccacg

47

<210> 8

<211> 49	
<212> DNA	
<213> HIV-1	
<400> 8	
ccggcgagcgggcggtcgagcgggcgtgagcgggcggatcgagcgggcg	49
<210> 9	
<211> 49	
<212> DNA	
<213> HIV-1	
<400> 9	
tcgacgcccgcgatccggcgtcacgcccgcgcaccgcccgcgcg	49
<210> 10	
<211> 50	
<212> DNA	
<213> HIV-1	
<400> 10	
ccggctgctgggcggcgtgcttgggcgtgcttgggcgggctgcttgggcg	50
<210> 11	
<211> 50	
<212> DNA	
<213> HIV-1	
<400> 11	
tcgacgcccagcagcccgcccaagcacgcccagcagccgcccagcag	50
<210> 12	
<211> 47	
<212> DNA	
<213> HIV-1	
<400> 12	
ccggactgggcgggggactgggcgtgactgggcggaggactgggcg	47
<210> 13	
<211> 47	
<212> DNA	
<213> HIV-1	
<400> 13	
tcgacgcccagtccttcgcccagtcacgcccagtccttcgcccagtc	47

<210> 14
 <211> 47
 <212> DNA
 <213> Homo sapiens

 <400> 14
 ccggagtgggcgggtggagtgggcgtgagtgggcggatggagtgggcg 47

 <210> 15
 <211> 47
 <212> DNA
 <213> Homo sapiens

 <400> 15
 tcgacgccactccatccgccactcacgccactccaccgccact 47

 <210> 16
 <211> 48
 <212> DNA
 <213> Homo sapiens

 <400> 16
 ccggacatgggcggagacatgggcgtacatgggcggaagacatgggcg 48

 <210> 17
 <211> 48
 <212> DNA
 <213> Homo sapiens

 <400> 17
 tcgacgccatgtcttcgccatgtacgccatgtctccgccatgt 48

 <210> 18
 <211> 120
 <212> DNA
 <213> 人工序列

 <220>
 <223> 质粒序列

 <221> CDS
 <222> (1)...(81)

 <400> 18
 aaa gag ggt ggg tcg acc ttc cgg act ggc cag gaa cgc cca gat ccg 48
 Lys Glu Gly Gly Ser Thr Phe Arg Thr Gly Gln Glu Arg Pro Asp Pro

1	5	10	15	
cgg gaa ttc aga tct act agt gcg gcc gct aag taagtaagac gtcgagctcg				101
Arg Glu Phe Arg Ser Thr Ser Ala Ala Ala Lys				
	20	25		
ccatcgcggtggaagcttt				120
<210> 19				
<211> 27				
<212> PRT				
<213> 人工序列				
<220>				
<223> 质粒序列				
<400> 19				
Lys Glu Gly Gly Ser Thr Phe Arg Thr Gly Gln Glu Arg Pro Asp Pro				
1	5	10	15	
Arg Glu Phe Arg Ser Thr Ser Ala Ala Ala Lys				
	20	25		
<210> 20				
<211> 303				
<212> DNA				
<213> 人工序列				
<220>				
<223> 质粒序列				
<221> CDS				
<222> (25)...(291)				
<400> 20				
gggtcgacct tccggactgg ccag gaa cgc cca tat gct tgc cct gtc gag				51
Glu Arg Pro Tyr Ala Cys Pro Val Glu				
	1	5		
tcc tgc gat cgc cgc ttt tct cgc tcg gat gag ctt acc cgc cat atc				99
Ser Cys Asp Arg Arg Phe Ser Arg Ser Asp Glu Leu Thr Arg His Ile				
10	15	20	25	
cgc atc cac act ggc cag aag ccc ttc cag tgt cga atc tgc atg cgt				147
Arg Ile His Thr Gly Gln Lys Pro Phe Gln Cys Arg Ile Cys Met Arg				
	30	35	40	

aac ttc agt cgt agt gac cac ctt acc acc cac atc cgg acc cac acc Asn Phe Ser Arg Ser Asp His Leu Thr Thr His Ile Arg Thr His Thr 45 50 55	195
ggc gag aag cct ttt gcc tgt gac att tgt ggg agg aag ttt gcc agg Gly Glu Lys Pro Phe Ala Cys Asp Ile Cys Gly Arg Lys Phe Ala Arg 60 65 70	243
agt gat gaa cgc aag agg cat acc aaa atc cat tta aga cag aag gat Ser Asp Glu Arg Lys Arg His Thr Lys Ile His Leu Arg Gln Lys Asp 75 80 85	291
ccgcgggaat cc	303
<210> 21 <211> 89 <212> PRT <213> 人工序列	
<220> <223> 质粒序列	
<400> 21 Glu Arg Pro Tyr Ala Cys Pro Val Glu Ser Cys Asp Arg Arg Phe Ser 1 5 10 15 Arg Ser Asp Glu Leu Thr Arg His Ile Arg Ile His Thr Gly Gln Lys 20 25 30 Pro Phe Gln Cys Arg Ile Cys Met Arg Asn Phe Ser Arg Ser Asp His 35 40 45 Leu Thr Thr His Ile Arg Thr His Thr Gly Glu Lys Pro Phe Ala Cys 50 55 60 Asp Ile Cys Gly Arg Lys Phe Ala Arg Ser Asp Glu Arg Lys Arg His 65 70 75 80 Thr Lys Ile His Leu Arg Gln Lys Asp 85	
<210> 22 <211> 69 <212> DNA <213> Homo sapiens	
<400> 22 tataaatgta agcaatgtgg gaaagctttt ggatgtcct caaaccttcg aaggcatgga aggactcac	60 69
<210> 23	

<211> 23

<212> PRT

<213> Homo sapiens

<400> 23

Tyr Lys Cys Lys Gln Cys Gly Lys Ala Phe Gly Cys Pro Ser Asn Leu

1

5

10

15

Arg Arg His Gly Arg Thr His

20

<210> 24

<211> 69

<212> DNA

<213> Homo sapiens

<400> 24

tataagtgtgta aggagtgtgg gaaagccttc aaccacagct ccaactcaa taaacaccac

60

agaatccac

69

<210> 25

<211> 23

<212> PRT

<213> Homo sapiens

<400> 25

Tyr Lys Cys Lys Glu Cys Gly Lys Ala Phe Asn His Ser Ser Asn Phe

1

5

10

15

Asn Lys His His Arg Ile His

20

<210> 26

<211> 69

<212> DNA

<213> Homo sapiens

<400> 26

tatgaatgta aggaatgtgg gaaagccttt agtagtggtt caaacttcac tcgacatcag

60

agaattcac

69

<210> 27

<211> 23

<212> PRT

<213> Homo sapiens

<400> 27

Tyr Glu Cys Lys Glu Cys Gly Lys Ala Phe Ser Ser Gly Ser Asn Phe

1

5

10

15

Thr Arg His Gln Arg Ile His
20

<210> 28

<211> 75

<212> DNA

<213> Homo sapiens

<400> 28

tatgtatgcg atgtagaggg atgtacgtgg aaatttggcc gctcagatga gctcaacaga	60
cacaagaaaaggcac	75

<210> 29

<211> 25

<212> PRT

<213> Homo sapiens

<400> 29

Tyr Val Cys Asp Val Glu Gly Cys Thr Trp Lys Phe Ala Arg Ser Asp	
1 5 10 15	
Glu Leu Asn Arg His Lys Lys Arg His	
20 25	

<210> 30

<211> 69

<212> DNA

<213> Homo sapiens

<400> 30

tatgagtgtatgaatgcgg gaaagctttt gcccaaaatt caactctcag agtacaccag	60
agaattcac	69

<210> 31

<211> 23

<212> PRT

<213> Homo sapiens

<400> 31

Tyr Glu Cys Asn Glu Cys Gly Lys Ala Phe Ala Gln Asn Ser Thr Leu	
1 5 10 15	
Arg Val His Gln Arg Ile His	
20	

<210> 32

<211> 69

<212> DNA

<213> Homo sapiens

<400> 32

tatgagtgtgta attactgtgg aaaaaccttt agtgtgagct caacccttat tagacatcag 60
 agaattccac 69

<210> 33

<211> 23

<212> PRT

<213> Homo sapiens

<400> 33

Tyr Glu Cys Asn Tyr Cys Gly Lys Thr Phe Ser Val Ser Ser Thr Leu
 1 5 10 15
 Ile Arg His Gln Arg Ile His
 20

<210> 34

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 34

tat cag tgc aac att tgc gga aaa tgt ttc tcc tgc aac tcc aac ctc 48
 Tyr Gln Cys Asn Ile Cys Gly Lys Cys Phe Ser Cys Asn Ser Asn Leu
 1 5 10 15

cac agg cac cag aga acg cac 69
 His Arg His Gln Arg Thr His
 20

<210> 35

<211> 23

<212> PRT

<213> Homo sapiens

<400> 35

Tyr Gln Cys Asn Ile Cys Gly Lys Cys Phe Ser Cys Asn Ser Asn Leu
 1 5 10 15
 His Arg His Gln Arg Thr His
 20

<210> 36

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 36

tat gca tgt cat cta tgt gga aaa gcc ttc act cag agt tct cac ctt 48

Tyr Ala Cys His Leu Cys Gly Lys Ala Phe Thr Gln Ser Ser His Leu

1 5 10 15

aga aga cat gag aaa act cac 69

Arg Arg His Glu Lys Thr His

20

<210> 37

<211> 23

<212> PRT

<213> Homo sapiens

<400> 37

Tyr Ala Cys His Leu Cys Gly Lys Ala Phe Thr Gln Ser Ser His Leu

1 5 10 15

Arg Arg His Glu Lys Thr His

20

<210> 38

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 38

tat aaa tgc ggc cag tgt ggg aag ttc tac tcg cag gtc tcc cac ctc 48

Tyr Lys Cys Gly Gln Cys Gly Lys Phe Tyr Ser Gln Val Ser His Leu

1 5 10 15

acc cgc cac cag aaa atc cac 69

Thr Arg His Gln Lys Ile His

20

<210> 39
 <211> 23
 <212> PRT
 <213> Homo sapiens

<400> 39
 Tyr Lys Cys Gly Gln Cys Gly Lys Phe Tyr Ser Gln Val Ser His Leu
 1 5 10 15
 Thr Arg His Gln Lys Ile His
 20

<210> 40
 <211> 69
 <212> DNA
 <213> Homo sapiens

<220>
 <221> CDS
 <222> (1)...(69)

<400> 40
 tat gca tgt cat cta tgt gga aaa gcc ttc act cag tgt tct cac ctt 48
 Tyr Ala Cys His Leu Cys Gly Lys Ala Phe Thr Gln Cys Ser His Leu
 1 5 10 15

aga aga cat gag aaa act cac 69
 Arg Arg His Glu Lys Thr His
 20

<210> 41
 <211> 23
 <212> PRT
 <213> Homo sapiens

<400> 41
 Tyr Ala Cys His Leu Cys Gly Lys Ala Phe Thr Gln Cys Ser His Leu
 1 5 10 15
 Arg Arg His Glu Lys Thr His
 20

<210> 42
 <211> 69
 <212> DNA
 <213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 42

tat gca tgt cat cta tgt gca aaa gcc ttc att cag tgt tct cac ctt 48

Tyr Ala Cys His Leu Cys Ala Lys Ala Phe Ile Gln Cys Ser His Leu

1

5

10

15

aga aga cat gag aaa act cac

69

Arg Arg His Glu Lys Thr His

20

<210> 43

<211> 23

<212> PRT

<213> Homo sapiens

<400> 43

Tyr Ala Cys His Leu Cys Ala Lys Ala Phe Ile Gln Cys Ser His Leu

1

5

10

15

Arg Arg His Glu Lys Thr His

20

<210> 44

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 44

tat gtt tgc agg gaa tgt ggg cgt ggc ttt cgc cag cat tca cac ctg 48

Tyr Val Cys Arg Glu Cys Gly Arg Gly Phe Arg Gln His Ser His Leu

1

5

10

15

gtc aga cac aag agg aca cat

69

Val Arg His Lys Arg Thr His

20

<210> 45

<211> 23

<212> PRT

<213> Homo sapiens

<400> 45

Tyr Val Cys Arg Glu Cys Gly Arg Gly Phe Arg Gln His Ser His Leu
 1 5 10 15
 Val Arg His Lys Arg Thr His
 20

<210> 46

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 46

ttt gag tgt aaa gat tgc ggg aaa gct ttc att cag aag tca aac ctc 48
 Phe Glu Cys Lys Asp Cys Gly Lys Ala Phe Ile Gln Lys Ser Asn Leu
 1 5 10 15

atc aga cac cag aga act cac 69
 Ile Arg His Gln Arg Thr His
 20

<210> 47

<211> 23

<212> PRT

<213> Homo sapiens

<400> 47

Phe Glu Cys Lys Asp Cys Gly Lys Ala Phe Ile Gln Lys Ser Asn Leu
 1 5 10 15
 Ile Arg His Gln Arg Thr His
 20

<210> 48

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 48

tat gtc tgc agg gag tgt agg cga ggt ttt agc cag aag tca aat ctc	48
Tyr Val Cys Arg Glu Cys Arg Arg Gly Phe Ser Gln Lys Ser Asn Leu	
1 5 10 15	

atc aga cac cag agg acg cac	69
Ile Arg His Gln Arg Thr His	
20	

<210> 49

<211> 23

<212> PRT

<213> Homo sapiens

<400> 49

Tyr Val Cys Arg Glu Cys Arg Arg Gly Phe Ser Gln Lys Ser Asn Leu	
1 5 10 15	
Ile Arg His Gln Arg Thr His	
20	

<210> 50

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 50

tat gaa tgt aac aca tgc agg aaa acc ttc tct caa aag tca aat ctc	48
Tyr Glu Cys Asn Thr Cys Arg Lys Thr Phe Ser Gln Lys Ser Asn Leu	
1 5 10 15	

att gta cat cag aga aca cac	69
Ile Val His Gln Arg Thr His	
20	

<210> 51

<211> 23

<212> PRT

<213> Homo sapiens

<400> 51

Tyr Glu Cys Asn Thr Cys Arg Lys Thr Phe Ser Gln Lys Ser Asn Leu
 1 5 10 15
 Ile Val His Gln Arg Thr His
 20

<210> 52
 <211> 69
 <212> DNA
 <213> Homo sapiens

<220>
 <221> CDS
 <222> (1)...(69)

<400> 52
 tat gtt tgc tca aaa tgt ggg aaa gcc ttc act cag agt tca aat ctg 48
 Tyr Val Cys Ser Lys Cys Gly Lys Ala Phe Thr Gln Ser Ser Asn Leu
 1 5 10 15

act gta cat caa aaa atc cac 69
 Thr Val His Gln Lys Ile His
 20

<210> 53
 <211> 23
 <212> PRT
 <213> Homo sapiens

<400> 53
 Tyr Val Cys Ser Lys Cys Gly Lys Ala Phe Thr Gln Ser Ser Asn Leu
 1 5 10 15
 Thr Val His Gln Lys Ile His
 20

<210> 54
 <211> 69
 <212> DNA
 <213> Homo sapiens

<220>
 <221> CDS
 <222> (1)...(69)

<400> 54
 tac aaa tgt gac gaa tgt gga aaa aac ttt acc cag tcc tcc aac ctt 48

Tyr Lys Cys Asp Glu Cys Gly Lys Asn Phe Thr Gln Ser Ser Asn Leu
 1 5 10 15

att gta cat aag aga att cat 69
 Ile Val His Lys Arg Ile His
 20

<210> 55

<211> 23

<212> PRT

<213> Homo sapiens

<400> 55

Tyr Lys Cys Asp Glu Cys Gly Lys Asn Phe Thr Gln Ser Ser Asn Leu
 1 5 10 15
 Ile Val His Lys Arg Ile His
 20

<210> 56

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 56

tat gaa tgt gat gtg tgt gga aaa acc ttc acg caa aag tca aac ctt 48
 Tyr Glu Cys Asp Val Cys Gly Lys Thr Phe Thr Gln Lys Ser Asn Leu
 1 5 10 15

ggt gta cat cag aga act cat 69
 Gly Val His Gln Arg Thr His
 20

<210> 57

<211> 23

<212> PRT

<213> Homo sapiens

<400> 57

Tyr Glu Cys Asp Val Cys Gly Lys Thr Phe Thr Gln Lys Ser Asn Leu
 1 5 10 15

Gly Val His Gln Arg Thr His
20

<210> 58

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 58

tat aag tgc cct gat tgt ggg aag agt ttt agt cag agt tcc agc ctc 48
Tyr Lys Cys Pro Asp Cys Gly Lys Ser Phe Ser Gln Ser Ser Ser Leu
1 5 10 15

att cgc cac cag cgg aca cac 69
Ile Arg His Gln Arg Thr His
20

<210> 59

<211> 23

<212> PRT

<213> Homo sapiens

<400> 59

Tyr Lys Cys Pro Asp Cys Gly Lys Ser Phe Ser Gln Ser Ser Ser Leu
1 5 10 15
Ile Arg His Gln Arg Thr His
20

<210> 60

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 60

tat gag tgt cag gac tgt ggg agg gcc ttc aac cag aac tcc tcc ctg 48
Tyr Glu Cys Gln Asp Cys Gly Arg Ala Phe Asn Gln Asn Ser Ser Leu
1 5 10 15

ggg cgg cac aag agg aca cac
Gly Arg His Lys Arg Thr His
20

69

<210> 61

<211> 23

<212> PRT

<213> Homo sapiens

<400> 61

Tyr Glu Cys Gln Asp Cys Gly Arg Ala Phe Asn Gln Asn Ser Ser Leu
1 5 10 15
Gly Arg His Lys Arg Thr His
20

<210> 62

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 62

tac aaa tgt gaa gaa tgt ggc aaa gct ttt aac cag tcc tca acc ctt 48
Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Asn Gln Ser Ser Thr Leu
1 5 10 15

act aga cat aag ata gtt cat
Thr Arg His Lys Ile Val His
20

69

<210> 63

<211> 23

<212> PRT

<213> Homo sapiens

<400> 63

Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Asn Gln Ser Ser Thr Leu
1 5 10 15
Thr Arg His Lys Ile Val His
20

<210> 64
 <211> 69
 <212> DNA
 <213> Homo sapiens

<220>
 <221> CDS
 <222> (1)...(69)

<400> 64
 tat aag tgc atg gag tgt ggg aag gct ttt aac cgc agg tca cac ctc 48
 Tyr Lys Cys Met Glu Cys Gly Lys Ala Phe Asn Arg Arg Ser His Leu
 1 5 10 15
 aca cgg cac cag cgg att cac 69
 Thr Arg His Gln Arg Ile His
 20

<210> 65
 <211> 23
 <212> PRT
 <213> Homo sapiens

<400> 65
 Tyr Lys Cys Met Glu Cys Gly Lys Ala Phe Asn Arg Arg Ser His Leu
 1 5 10 15
 Thr Arg His Gln Arg Ile His
 20

<210> 66
 <211> 69
 <212> DNA
 <213> Homo sapiens

<220>
 <221> CDS
 <222> (1)...(69)

<400> 66
 tat aca tgt aaa cag tgt ggg aaa gcc ttc agt gtt tcc agt tcc ctt 48
 Tyr Thr Cys Lys Gln Cys Gly Lys Ala Phe Ser Val Ser Ser Ser Leu
 1 5 10 15
 cga aga cat gaa acc act cac 69

Arg Arg His Glu Thr Thr His
20

<210> 67

<211> 23

<212> PRT

<213> Homo sapiens

<400> 67

Tyr Thr Cys Lys Gln Cys Gly Lys Ala Phe Ser Val Ser Ser Ser Leu
1 5 10 15
Arg Arg His Glu Thr Thr His
20

<210> 68

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 68

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa
1 5 10 15
Ser Asn Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His
20 25

<210> 69

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 69

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	His	Xaa
1			5				10							15	
Ser	Asn	Xaa	Xaa	Lys	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
			20				25								

<210> 70

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 70

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Ser	Xaa
1			5				10							15	
Ser	Asn	Xaa	Xaa	Arg	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				

20 <210> 71 <211> 28 <212> PRT <213> 人工序列 <220> <223> 纯化多肽 <221> VARIANT <222> 2, 4-8, 10-12, 14, 16, 20, 23-27 <223> Xaa = 任何氨基酸 <221> VARIANT <222> 1, 13 <223> Xaa = Phe或Tyr <221> VARIANT <222> 19 <223> Xaa = 疏水性残基 <400> 71 Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Gln Xaa 1 5 10 15 Ser Thr Xaa Xaa Val His Xaa Xaa Xaa Xaa Xaa His 20 25 <210> 72 <211> 28 <212> PRT <213> 人工序列 <220> <223> 纯化多肽 <221> VARIANT <222> 2, 4-8, 10-12, 14, 16, 20, 23-27 <223> Xaa = 任何氨基酸 <221> VARIANT <222> 1, 13 <223> Xaa = Phe或Tyr <221> VARIANT	25
--	-----------

<222> 18

<223> Xaa = Ser或Thr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 72

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Val Xaa

1

5

10

15

Ser Xaa Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His

20

25

<210> 73

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 73

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Gln Xaa

1

5

10

15

Ser His Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His

20

25

<210> 74

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 74

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Gln	Xaa
1				5				10						15	
Ser	Asn	Xaa	Xaa	Val	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
				20				25							

<210> 75

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 18

<223> Xaa = Ser或Thr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 75

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Gln Xaa
 1 5 10 15
 Ser Xaa Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His
 20 25

<210> 76
 <211> 28
 <212> PRT
 <213> 人工序列

<220>
 <223> 协调残基

<221> VARIANT
 <222> 1, 13
 <223> Xaa = Phe或Tyr

<221> VARIANT
 <222> 2, 4-8, 10-12, 14, 16, 20, 23-27
 <223> Xaa = 任何氨基酸

<221> VARIANT
 <222> 19
 <223> Xaa = 疏水性残基

<400> 76
 Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Xaa Xaa
 1 5 10 15
 Xaa Xaa Xaa Xaa Xaa His Xaa Xaa Xaa Xaa Xaa His
 20 25

<210> 77
 <211> 24
 <212> PRT
 <213> 人工序列

<220>
 <223> 多肽基序

<221> VARIANT
 <222> 1
 <223> Xaa = Leu, Ile, Val, Met, Phe, Tyr,或Gly

<221> VARIANT
 <222> 2

<223> Xaa = Ala, Ser, Leu, Val,或Arg

<221> VARIANT

<222> 3-4, 6, 8-11, 17, 19-23

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 5

<223> Xaa = Leu, Ile, Val, Met, Ser, Thr, Ala, Cys,或Asn

<221> VARIANT

<222> 7

<223> Xaa = Leu, Ile, Val,或Met

<221> VARIANT

<222> 12

<223> Xaa = Leu, Ile,或Val

<221> VARIANT

<222> 13

<223> Xaa = Arg, Lys, Asn, Gln, Glu, Ser, Thr, Ala, Ile,或Tyr

<221> VARIANT

<222> 14

<223> Xaa = Leu, Ile, Val, Phe, Ser, Thr, Asn, Lys, or
His

<221> VARIANT

<222> 16

<223> Xaa = Phe, Tyr, Val,或Cys

<221> VARIANT

<222> 18

<223> Xaa = Asn, Asp, Gln, Thr, Ala,或His

<221> VARIANT

<222> 24

<223> Xaa = Arg, Lys, Asn, Ala, Ile, Met,或Trp

<400> 77

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Trp Xaa

1

5

10

15

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa

20

<210> 78
 <211> 6
 <212> PRT
 <213> Eukaryote

<220>
 <221> VARIANT
 <222> 3
 <223> Xaa = Glu或Gln

<221> VARIANT
 <222> 4
 <223> Xaa = Lys或Arg

<221> VARIANT
 <222> 6
 <223> Xaa = Tyr或Phe

<400> 78
 Thr Gly Xaa Xaa Pro Xaa
 1 5

<210> 79
 <211> 29
 <212> DNA
 <213> 人工序列

<220>
 <223> 合成寡核苷酸

<400> 79
 tgcctgcagc atttgtggga ggaagtttg

29

<210> 80
 <211> 30
 <212> DNA
 <213> 人工序列

<220>
 <223> 合成寡核苷酸

<400> 80
 atgctgcagg ctaaggctt ctcgccggtg

30

<210> 81
<211> 24
<212> DNA
<213> 人工序列

<220>
<223> PCR引物

<221> misc_feature
<222> 11, 17, 20
<223> n = A, T, G,或C;

<400> 81
gcgtccggac ncayacnggn sara

24

<210> 82
<211> 24
<212> DNA
<213> 人工序列

<220>
<223> PCR引物

<221> misc_feature
<222> 10-11, 16,
<223> n = A, T, G,或C;

<400> 82
cggaattcan nbrwanggyy tytc

24

<210> 83
<211> 7
<212> PRT
<213> 人工序列

<220>
<223> 氨基酸基序

<221> VARIANT
<222> 4
<223> Xaa = Glu或Gln

<221> VARIANT

<222> 5

<223> Xaa = Lys或Arg

<221> VARIANT

<222> 3

<223> Xaa = Tyr或Phe

<400> 83

His Thr Gly Xaa Xaa Pro Xaa

1

5

<210> 84

<211> 54

<212> DNA

<213> 人工序列

<220>

<223> 合成寡核苷酸

<400> 84

gggcccgggg agaagcctta cgcattgtcca gtcgaatctt gtgatagaag attc

54

<210> 85

<211> 75

<212> DNA

<213> 人工序列

<220>

<223> 合成寡核苷酸

<221> misc_feature

<222> 36, 39, 45, 51, 54,

<223> n = A, T, G,或C;

<400> 85

ctccccgcgg ttgcgcgtg tggattctga tatgsnbsnb aagsnbsnbs nbsnbtgaga

60

atcttctatc acaag

75

<210> 86

<211> 23

<212> DNA

<213> 人工序列

<220>

<223> 合成寡核苷酸

<400> 86

ctagacccgg gaattcgtcg acg

23

<210> 87

<211> 23

<212> DNA

<213> 人工序列

<220>

<223> 合成寡核苷酸

<400> 87

gatccgtcga cgaattcccg ggt

23

<210> 88

<211> 38

<212> DNA

<213> 人工序列

<220>

<223> 合成寡核苷酸

<221> misc_feature

<222> 6-8, 18-20, 30-32

<223> n = A, T, G,或C

<400> 88

ccggtnnntg ggcgtacnnn tgggcgtcan nntgggcg

38

<210> 89

<211> 38

<212> DNA

<213> 人工序列

<220>

<223> 合成寡核苷酸

<221> misc_feature

<222> 11-13, 23-25, 35-37

<223> n = A, T, G,或C

<400> 89 tcgacgcccc nnntgacgcc cannngtacg cccannna	38
<210> 90 <211> 24 <212> DNA <213> 人工序列	
<220> <223> 用于凝胶移位分析的合成探针	
<400> 90 ccgggtcgcg cgtgggcggt accg	24
<210> 91 <211> 24 <212> DNA <213> 人工序列	
<220> <223> 用于凝胶移位分析的合成探针	
<400> 91 tcgacggtac cgcccacgcg cgac	24
<210> 92 <211> 24 <212> DNA <213> 人工序列	
<220> <223> 用于凝胶移位分析的合成探针	
<400> 92 ccgggtcgcg agcgggcggt accg	24
<210> 93 <211> 24 <212> DNA <213> 人工序列	

<220>

<223> 用于凝胶移位分析的合成探针

<400> 93

tcgacggtac cgcccgtcg cgac

24

<210> 94

<211> 24

<212> DNA

<213> 人工序列

<220>

<223> 用于凝胶移位分析的合成探针

<400> 94

ccgggtcgtg cttgggcggt accg

24

<210> 95

<211> 24

<212> DNA

<213> 人工序列

<220>

<223> 用于凝胶移位分析的合成探针

<400> 95

tcgacggtac cgcccaagca cgac

24

<210> 96

<211> 24

<212> DNA

<213> 人工序列

<220>

<223> 用于凝胶移位分析的合成探针

<400> 96

ccgggtcggg actgggcggt accg

24

<210> 97

<211> 24

<212> DNA

<213> 人工序列

<220>

<223> 用于凝胶移位分析的合成探针

<400> 97

tcgacggtac cgcccagtcg cgac

24

<210> 98

<211> 24

<212> DNA

<213> 人工序列

<220>

<223> 用于凝胶移位分析的合成探针

<400> 98

ccgggtcggg agtgggcggt accg

24

<210> 99

<211> 24

<212> DNA

<213> 人工序列

<220>

<223> 用于凝胶移位分析的合成探针

<400> 99

tcgacggtac cgcccactcc cgac

24

<210> 100

<211> 24

<212> DNA

<213> 人工序列

<220>

<223> 用于凝胶移位分析的合成探针

<400> 100

ccgggtcggg catgggcggt accg

24

<210> 101

<211> 24

<212> DNA

<213> 人工序列

<220>

<223> 用于凝胶移位分析的合成探针

<400> 101

tcgacggtac cgcccatgtc cgac

24

<210> 102

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 102

tat aag tgt aag gaa tgt ggg cag gcc ttt aga cag cgt gca cat ctt

48

Tyr Lys Cys Lys Glu Cys Gly Gln Ala Phe Arg Gln Arg Ala His Leu

1

5

10

15

att cga cat cac aaa ctt cac

69

Ile Arg His His Lys Leu His

20

<210> 103

<211> 23

<212> PRT

<213> Homo sapiens

<400> 103

Tyr Lys Cys Lys Glu Cys Gly Gln Ala Phe Arg Gln Arg Ala His Leu

1

5

10

15

Ile Arg His His Lys Leu His

20

<210> 104

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 104

tat aag tgt cat caa tgt ggg aaa gcc ttt att caa tcc ttt aac ctt 48

Tyr Lys Cys His Gln Cys Gly Lys Ala Phe Ile Gln Ser Phe Asn Leu

1 5 10 15

cga aga cat gag aga act cac

69

Arg Arg His Glu Arg Thr His

20

<210> 105

<211> 23

<212> PRT

<213> Homo sapiens

<400> 105

Tyr Lys Cys His Gln Cys Gly Lys Ala Phe Ile Gln Ser Phe Asn Leu

1 5 10 15

Arg Arg His Glu Arg Thr His

20

<210> 106

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 106

ttc cag tgt aat cag tgt ggg gca tct ttt act cag aaa ggt aac ctc 48

Phe Gln Cys Asn Gln Cys Gly Ala Ser Phe Thr Gln Lys Gly Asn Leu

1 5 10 15

ctc cgc cac att aaa ctg cac

69

Leu Arg His Ile Lys Leu His

20

<210> 107

<211> 23

<212> PRT

<213> Homo sapiens

<400> 107

Phe Gln Cys Asn Gln Cys Gly Ala Ser Phe Thr Gln Lys Gly Asn Leu

1

5

10

15

Leu Arg His Ile Lys Leu His

20

<210> 108

<211> 72

<212> DNA

<213> 人工序列

<220>

<223> PCR引物

<221> misc_feature

<222> 22-72

<223> n = A, T, G,或C

<400> 108

accacactg gccagaaacc cnnnnnnnnnn nnnnnnnnnnn nnnnnnnnnnn nnnnnnnnnnn 60

nnnnnnnnnn nn 72

<210> 109

<211> 66

<212> DNA

<213> 人工序列

<220>

<223> PCR引物

<221> misc_feature

<222> 22-66

<223> n = A, T, G,或C

<400> 109

gatctgaatt cattcaccgg tnnnnnnnnnn nnnnnnnnnnn nnnnnnnnnnn nnnnnnnnnnn 60

nnnnnnn 66

<210> 110

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 110

tac aaa tgt gaa gaa tgt ggc aaa gcc ttt agg cag tcc tca cac ctt 48

Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Arg Gln Ser Ser His Leu

1

5

10

15

act aca cat aag ata att cat 69

Thr Thr His Lys Ile Ile His

20

<210> 111

<211> 23

<212> PRT

<213> Homo sapiens

<400> 111

Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Arg Gln Ser Ser His Leu

1

5

10

15

Thr Thr His Lys Ile Ile His

20

<210> 112

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 112

tat gag tgt gat cac tgt gga aaa tcc ttt agc cag agc tct cat ctg 48

Tyr Glu Cys Asp His Cys Gly Lys Ser Phe Ser Gln Ser Ser His Leu

1

5

10

15

aat gtg cac aaa aga act cac 69

Asn Val His Lys Arg Thr His

20

<210> 113

<211> 23

<212> PRT

<213> Homo sapiens

<400> 113

Tyr Glu Cys Asp His Cys Gly Lys Ser Phe Ser Gln Ser Ser His Leu

1

5

10

15

Asn Val His Lys Arg Thr His

20

<210> 114

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 114

tac atg tgc agt gag tgt ggg cga ggc ttc agc cag aag tca aac ctc 48

Tyr Met Cys Ser Glu Cys Gly Arg Gly Phe Ser Gln Lys Ser Asn Leu

1

5

10

15

atc ata cac cag agg aca cac 69

Ile Ile His Gln Arg Thr His

20

<210> 115

<211> 23

<212> PRT

<213> Homo sapiens

<400> 115

Tyr Met Cys Ser Glu Cys Gly Arg Gly Phe Ser Gln Lys Ser Asn Leu

1

5

10

15

Ile Ile His Gln Arg Thr His

20

<210> 116

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 116

tat gaa tgt gaa aaa tgt ggc aaa gct ttt aac cag tcc tca aat ctt 48

Tyr Glu Cys Glu Lys Cys Gly Lys Ala Phe Asn Gln Ser Ser Asn Leu

1 5 10 15

act aga cat aag aaa agt cat 69

Thr Arg His Lys Lys Ser His

20

<210> 117

<211> 23

<212> PRT

<213> Homo sapiens

<400> 117

Tyr Glu Cys Glu Lys Cys Gly Lys Ala Phe Asn Gln Ser Ser Asn Leu

1 5 10 15

Thr Arg His Lys Lys Ser His

20

<210> 118

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 118

tat gag tgc aat gaa tgt ggg aag ttt ttt agc cag agc tcc agc ctc 48

Tyr Glu Cys Asn Glu Cys Gly Lys Phe Phe Ser Gln Ser Ser Ser Leu

1 5 10 15

att aga cat agg aga agt cac 69

Ile Arg His Arg Arg Ser His

20

<210> 119

<211> 23

<212> PRT

<213> Homo sapiens

<400> 119

Tyr Glu Cys Asn Glu Cys Gly Lys Phe Phe Ser Gln Ser Ser Ser Leu

1

5

10

15

Ile Arg His Arg Arg Ser His

20

<210> 120

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 120

tat gag tgt cac gat tgc gga aag tcc ttt agg cag agc acc cac ctc

48

Tyr Glu Cys His Asp Cys Gly Lys Ser Phe Arg Gln Ser Thr His Leu

1

5

10

15

act cag cac cgg agg atc cac

69

Thr Gln His Arg Arg Ile His

20

<210> 121

<211> 23

<212> PRT

<213> Homo sapiens

<400> 121

Tyr Glu Cys His Asp Cys Gly Lys Ser Phe Arg Gln Ser Thr His Leu

1

5

10

15

Thr Gln His Arg Arg Ile His

20

<210> 122

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 122

tat gag tgt cac gat tgc gga aag tcc ttt agg cag agc acc cac ctc 48

Tyr Glu Cys His Asp Cys Gly Lys Ser Phe Arg Gln Ser Thr His Leu

1

5

10

15

act cgg cac cgg agg atc cac

69

Thr Arg His Arg Arg Ile His

20

<210> 123

<211> 23

<212> PRT

<213> Homo sapiens

<400> 123

Tyr Glu Cys His Asp Cys Gly Lys Ser Phe Arg Gln Ser Thr His Leu

1

5

10

15

Thr Arg His Arg Arg Ile His

20

<210> 124

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 124

cac aag tgc ctt gaa tgt ggg aaa tgc ttc agt cag aac acc cat ctg 48

His Lys Cys Leu Glu Cys Gly Lys Cys Phe Ser Gln Asn Thr His Leu

1

5

10

15

act cgc cac caa cgc acc cac

69

Thr Arg His Gln Arg Thr His

20

<210> 125

<211> 23

<212> PRT

<213> Homo sapiens

<400> 125

His Lys Cys Leu Glu Cys Gly Lys Cys Phe Ser Gln Asn Thr His Leu

1

5

10

15

Thr Arg His Gln Arg Thr His

20

<210> 126

<211> 75

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(75)

<400> 126

tac cac tgt gac tgg gac ggc tgt gga tgg aaa ttc gcc cgc tca gat

48

Tyr His Cys Asp Trp Asp Gly Cys Gly Trp Lys Phe Ala Arg Ser Asp

1

5

10

15

gaa ctg acc agg cac tac cgt aaa cac

75

Glu Leu Thr Arg His Tyr Arg Lys His

20

25

<210> 127

<211> 25

<212> PRT

<213> Homo sapiens

<400> 127

Tyr His Cys Asp Trp Asp Gly Cys Gly Trp Lys Phe Ala Arg Ser Asp

1

5

10

15

Glu Leu Thr Arg His Tyr Arg Lys His

20

25

<210> 128

<211> 75

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(75)

<400> 128

tac aga tgc tca tgg gaa ggg tgt gag tgg cgt ttt gca aga agt gat 48

Tyr Arg Cys Ser Trp Glu Gly Cys Glu Trp Arg Phe Ala Arg Ser Asp

1 5 10 15

gag tta acc agg cac ttc cga aag cac 75

Glu Leu Thr Arg His Phe Arg Lys His

20 25

<210> 129

<211> 25

<212> PRT

<213> Homo sapiens

<400> 129

Tyr Arg Cys Ser Trp Glu Gly Cys Glu Trp Arg Phe Ala Arg Ser Asp

1 5 10 15

Glu Leu Thr Arg His Phe Arg Lys His

20 25

<210> 130

<211> 75

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(75)

<400> 130

ttc agc tgt agc tgg aaa ggt tgt gaa agg agg ttt gcc cgt tct gat 48

Phe Ser Cys Ser Trp Lys Gly Cys Glu Arg Arg Phe Ala Arg Ser Asp

1 5 10 15

gaa ctg tcc aga cac agg cga acc cac 75

Glu Leu Ser Arg His Arg Arg Thr His

20 25

<210> 131

<211> 25

<212> PRT

<213> Homo sapiens

<400> 131

Phe Ser Cys Ser Trp Lys Gly Cys Glu Arg Arg Phe Ala Arg Ser Asp
 1 5 10 15
 Glu Leu Ser Arg His Arg Arg Thr His
 20 25

<210> 132

<211> 75

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(75)

<400> 132

ttc gcc tgc agc tgg cag gac tgc aac aag aag ttc gcg cgc tcc gac 48
 Phe Ala Cys Ser Trp Gln Asp Cys Asn Lys Lys Phe Ala Arg Ser Asp
 1 5 10 15

gag ctg gcg cgg cac tac cgc aca cac 75
 Glu Leu Ala Arg His Tyr Arg Thr His
 20 25

<210> 133

<211> 25

<212> PRT

<213> Homo sapiens

<400> 133

Phe Ala Cys Ser Trp Gln Asp Cys Asn Lys Lys Phe Ala Arg Ser Asp
 1 5 10 15
 Glu Leu Ala Arg His Tyr Arg Thr His
 20 25

<210> 134

<211> 75

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(75)

<400> 134

tac cac tgc aac tgg gac ggc tgc ggc tgg aag ttt gcg cgc tca gac 48
Tyr His Cys Asn Trp Asp Gly Cys Gly Trp Lys Phe Ala Arg Ser Asp
1 5 10 15

gag ctc acg cgc cac tac cga aag cac 75
Glu Leu Thr Arg His Tyr Arg Lys His
20 25

<210> 135

<211> 25

<212> PRT

<213> Homo sapiens

<400> 135

Tyr His Cys Asn Trp Asp Gly Cys Gly Trp Lys Phe Ala Arg Ser Asp
1 5 10 15
Glu Leu Thr Arg His Tyr Arg Lys His
20 25

<210> 136

<211> 72

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(72)

<400> 136

ttc ctc tgt cag tat tgt gca cag aga ttt ggg cga aag gat cac ctg 48
Phe Leu Cys Gln Tyr Cys Ala Gln Arg Phe Gly Arg Lys Asp His Leu
1 5 10 15

act cga cat atg aag aag agt cac 72
Thr Arg His Met Lys Lys Ser His
20

<210> 137

<211> 24

<212> PRT

<213> Homo sapiens

<400> 137

Phe Leu Cys Gln Tyr Cys Ala Gln Arg Phe Gly Arg Lys Asp His Leu

1

5

10

15

Thr Arg His Met Lys Lys Ser His

20

<210> 138

<211> 78

<212> DNA

<213> 人工序列

<220>

<223> PCR引物

<400> 138

tgtcgaatct gcatgcgtaa ctfcagtcgt agtgaccacc ttaccacca catccggacc

60

cacactggcc agaaaccc

78

<210> 139

<211> 81

<212> DNA

<213> 人工序列

<220>

<223> PCR引物

<400> 139

ggtggcggcc gttacttact tagagctcga cgtcttactt acttagcggc cgcactagta

60

gatctgaatt cattcaccgg t

81

<210> 140

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 140

ttc cag tgt aaa act tgt cag cga aag ttc tcc cgg tcc gac cac ctg

48

Phe Gln Cys Lys Thr Cys Gln Arg Lys Phe Ser Arg Ser Asp His Leu

1

5

10

15

aag acc cac acc agg act cat
Lys Thr His Thr Arg Thr His
20

69

<210> 141
<211> 23
<212> PRT
<213> Homo sapiens

<400> 141
Phe Gln Cys Lys Thr Cys Gln Arg Lys Phe Ser Arg Ser Asp His Leu
1 5 10 15
Lys Thr His Thr Arg Thr His
20

<210> 142
<211> 69
<212> DNA
<213> Homo sapiens

<220>
<221> CDS
<222> (1)...(69)

<400> 142
ttt gcc tgc gag gtc tgc ggt gtt cga ttc acc agg aac gac aag ctg
Phe Ala Cys Glu Val Cys Gly Val Arg Phe Thr Arg Asn Asp Lys Leu
1 5 10 15

48

aag atc cac atg cgg aag cac
Lys Ile His Met Arg Lys His
20

69

<210> 143
<211> 23
<212> PRT
<213> Homo sapiens

<400> 143
Phe Ala Cys Glu Val Cys Gly Val Arg Phe Thr Arg Asn Asp Lys Leu
1 5 10 15
Lys Ile His Met Arg Lys His
20

<210> 144

<211> 75

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(75)

<400> 144

tat gta tgc gat gta gag gga tgt acg tgg aaa ttt gcc cgc tca gat	48
Tyr Val Cys Asp Val Glu Gly Cys Thr Trp Lys Phe Ala Arg Ser Asp	
1 5 10 15	

aag ctc aac aga cac aag aaa agg cac	75
Lys Leu Asn Arg His Lys Lys Arg His	
20 25	

<210> 145

<211> 25

<212> PRT

<213> Homo sapiens

<400> 145

Tyr Val Cys Asp Val Glu Gly Cys Thr Trp Lys Phe Ala Arg Ser Asp	
1 5 10 15	
Lys Leu Asn Arg His Lys Lys Arg His	
20 25	

<210> 146

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 146

tat att tgc aga aag tgt gga cgg ggc ttt agt cgg aag tcc aac ctt	48
Tyr Ile Cys Arg Lys Cys Gly Arg Gly Phe Ser Arg Lys Ser Asn Leu	
1 5 10 15	

atc aga cat cag agg aca cac	69
-----------------------------	----

Ile Arg His Gln Arg Thr His
20

<210> 147

<211> 23

<212> PRT

<213> Homo sapiens

<400> 147

Tyr Ile Cys Arg Lys Cys Gly Arg Gly Phe Ser Arg Lys Ser Asn Leu
1 5 10 15

Ile Arg His Gln Arg Thr His
20

<210> 148

<211> 69

<212> DNA

<213> Homo sapiens

<220>

<221> CDS

<222> (1)...(69)

<400> 148

tat cta tgt agt gag tgt gac aaa tgc ttc agt aga agt aca aac ctc 48

Tyr Leu Cys Ser Glu Cys Asp Lys Cys Phe Ser Arg Ser Thr Asn Leu
1 5 10 15

ata agg cat cga aga act cac 69
Ile Arg His Arg Arg Thr His
20

<210> 149

<211> 23

<212> PRT

<213> Homo sapiens

<400> 149

Tyr Leu Cys Ser Glu Cys Asp Lys Cys Phe Ser Arg Ser Thr Asn Leu
1 5 10 15

Ile Arg His Arg Arg Thr His
20

<210> 150

<223> Xaa = 疏水性残基

<400> 151

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Gln	Xaa
1				5				10						15	
Phe	Asn	Xaa	Xaa	Arg	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
			20					25							

<210> 152

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 152

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Gln	Xaa
1				5				10						15	
Ser	His	Xaa	Xaa	Thr	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
			20					25							

<210> 153

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 153

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Gln Xaa

1 5 10 15

Ser His Xaa Xaa Val His Xaa Xaa Xaa Xaa Xaa His

20 25

<210> 154

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 154

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Gln Xaa

1 5 10 15

Ser Asn Xaa Xaa Ile His Xaa Xaa Xaa Xaa Xaa His

20 25

<210> 155

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 155

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Gln	Xaa
1				5				10						15	
Ser	Asn	Xaa	Xaa	Arg	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
			20				25								

<210> 156

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 156

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Gln	Xaa
1				5				10						15	
Thr	His	Xaa	Xaa	Gln	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
			20				25								

<210> 157

<211> 26

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 2-6, 12, 14, 18, 21-26

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 11

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 17

<223> Xaa = 疏水性残基

<400> 157

Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Gln Xaa Thr His

1

5

10

15

Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His

20

25

<210> 158

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 158

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Arg	Xaa
1				5				10						15	
Asp	Lys	Xaa	Xaa	Ile	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
				20				25							

<210> 159

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 159

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Arg	Xaa
1				5				10						15	
Ser	Asn	Xaa	Xaa	Arg	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
				20				25							

<210> 160

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 160

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Arg Xaa

1

5

10

15

Thr Asn Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His

20

25

<210> 161

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8 ,10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 161

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Gln Xaa

1

5

10

15

Gly Asn Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His

20

25

<210> 162

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 162

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Arg Xaa

1

5

10

15

Asp Glu Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His

20

25

<210> 163

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 163

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Arg Xaa

1

5

10

15

Asp His Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His
 20 25

<210> 164

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 164

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Arg Xaa
 1 5 10 15
 Asp His Xaa Xaa Thr His Xaa Xaa Xaa Xaa Xaa His
 20 25

<210> 165

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 165

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Arg Xaa

1

5

10

15

Asp Lys Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His

20

25

<210> 166

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27)

<223> Xaa = 任何氨基酸

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<400> 166

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Arg Xaa

1

5

10

15

Ser His Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His

20

25

<210> 167

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<400> 167

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	His	Xaa
1			5					10						15	
Ser	Ser	Xaa	Xaa	Arg	His	Xaa	Xaa	Xaa	Xaa	Xaa	Xaa	His			
			20				25								

<210> 168

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<400> 168

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Ile	Xaa
1				5					10					15	
Ser	Asn	Xaa	Xaa	Arg	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
			20				25								

<210> 169

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<221> VARIANT

<222> 2, 4 -8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<400> 169

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Lys Xaa

1

5

10

15

Ser Asn Xaa Xaa Arg His Xaa Xaa Xaa Xaa Xaa His

20

25

<210> 170

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<400> 170

Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Gln Xaa
 1 5 10 15
 Ser Asn Xaa Xaa Lys His Xaa Xaa Xaa Xaa Xaa His
 20 25

<210> 171
 <211> 28
 <212> PRT
 <213> 人工序列

<220>
 <223> 纯化多肽

<221> VARIANT
 <222> 1, 13
 <223> Xaa = Phe或Tyr

<221> VARIANT
 <222> 19
 <223> Xaa = 疏水性残基

<221> VARIANT
 <222> 2, 4-8, 10-12, 14, 16, 20, 23-27
 <223> Xaa = 任何氨基酸

<400> 171
 Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Cys Xaa Xaa Xaa Xaa Xaa Gln Xaa
 1 5 10 15
 Ser His Xaa Xaa Thr His Xaa Xaa Xaa Xaa Xaa His
 20 25

<210> 172
 <211> 28
 <212> PRT
 <213> 人工序列

<220>
 <223> 纯化多肽

<221> VARIANT
 <222> 1, 13
 <223> Xaa = Phe或Tyr

<221> VARIANT
 <222> 19

<223> Xaa = 疏水性残基

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<400> 172

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Val	Xaa
1			5					10						15	
Ser	Asn	Xaa	Xaa	Val	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
			20					25							

<210> 173

<211> 23

<212> PRT

<213> Homo sapiens

<400> 173

Phe	Lys	Cys	Pro	Val	Cys	Gly	Lys	Ala	Phe	Arg	His	Ser	Ser	Ser	Leu
1			5					10						15	
Val	Arg	His	Gln	Arg	Thr	His									
			20												

<210> 174

<211> 69

<212> DNA

<213> Homo sapiens

<400> 174

ttcaagtgcc	cagtgtgcgg	caaggccttc	cggcatagct	cctcgctggt	gcggcaccag	60
cgcacgcac						69

<210> 175

<211> 24

<212> PRT

<213> Homo sapiens

<400> 175

Tyr	Arg	Cys	Lys	Tyr	Cys	Asp	Arg	Ser	Phe	Ser	Ile	Ser	Ser	Asn	Leu
1			5					10						15	
Gln	Arg	His	Val	Arg	Asn	Ile	His								
			20												

<210> 176

<211> 72

184

aaaattcat

69

<210> 181

<211> 23

<212> PRT

<213> Homo sapiens

<400> 181

Tyr Glu Cys Val Gln Cys Gly Lys Gly Phe Thr Gln Ser Ser Asn Leu

1

5

10

15

Ile Thr His Gln Arg Val His

20

<210> 182

<211> 69

<212> DNA

<213> Homo sapiens

<400> 182

tacgagtgtg tgcagtgtgg gaaaggttc acccagagct ccaacctcat cacacatcaa

agagttcac

60

69

<210> 183

<211> 23

<212> PRT

<213> Homo sapiens

<400> 183

Tyr Glu Cys Asp His Cys Gly Lys Ala Phe Ser Val Ser Ser Asn Leu

1

5

10

15

Asn Val His Arg Arg Ile His

20

<210> 184

<211> 69

<212> DNA

<213> Homo sapiens

<400> 184

tatgaatgcg atcactgtgg gaaagccttc agcgtcagct ccaacctgaa cgtgcacaga

aggatccac

60

69

<210> 185

<211> 23

<212> PRT

<213> Homo sapiens

<400> 185

Tyr Thr Cys Ser Asp Cys Gly Lys Ala Phe Arg Asp Lys Ser Cys Leu

1

5

10

15

Asn Arg His Arg Arg Thr His

20

<210> 186

<211> 69

<212> DNA

<213> Homo sapiens

<400> 186

tacacatgca	gtgactgtgg	gaaagcttc	agagataaat	catgtctcaa	cagacatcgg	60
agaactcat						69

<210> 187

<211> 23

<212> PRT

<213> Homo sapiens

<400> 187

Tyr Glu Cys Asp His Cys Gly Lys Ala Phe Ser Ile Gly Ser Asn Leu

1

5

10

15

Asn Val His Arg Arg Ile His

20

<210> 188

<211> 69

<212> DNA

<213> Homo sapiens

<400> 188

tacgaatgcg	atcactgtgg	gaaggccttc	agcataggct	ccaacctgaa	tgtgcacagg	60
cggatccat						69

<210> 189

<211> 23

<212> PRT

<213> Homo sapiens

<400> 189

Tyr Arg Cys Glu Glu Cys Gly Lys Ala Phe Arg Trp Pro Ser Asn Leu

187

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<400> 193

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Asp	Xaa
1				5				10						15	
Ser	Cys	Xaa	Xaa	Arg	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
			20				25								

<210> 194

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<400> 194

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Ile	Xaa
1				5				10						15	
Ser	Asn	Xaa	Xaa	Val	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
			20				25								

<210> 195

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<400> 195

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Trp	Xaa
1			5				10							15	
Ser	Asn	Xaa	Xaa	Arg	His	Xaa	Xaa	Xaa	Xaa	Xaa	His				
			20				25								

<210> 196

<211> 28

<212> PRT

<213> 人工序列

<220>

<223> 纯化多肽

<221> VARIANT

<222> 1, 13

<223> Xaa = Phe或Tyr

<221> VARIANT

<222> 19

<223> Xaa = 疏水性残基

<221> VARIANT

<222> 2, 4-8, 10-12, 14, 16, 20, 23-27

<223> Xaa = 任何氨基酸

<400> 196

Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Cys	Xaa	Xaa	Xaa	Xaa	Xaa	Asp	Xaa
1			5				10							15	

Ser Ala Xaa Xaa Arg His Xaa Xaa Xaa Xaa His
20 25

<210> 197

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 197

gcgtgggcgt

10

<210> 198

<211> 56

<212> PRT

<213> Homo sapiens

<400> 198

Glu Arg Pro Tyr Ala Cys Pro Val Glu Ser Cys Asp Arg Arg Phe Ser

1

5

10

15

Arg Ser Asp Glu Leu Thr Arg His Ile Arg Ile His Thr Gly Gln Lys

20

25

30

Pro Phe Gln Cys Arg Ile Cys Met Arg Asn Phe Ser Arg Ser Asp His

35

40

45

Leu Thr Thr His Ile Arg Thr His

50

55

<210> 199

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 199

gagagggagc

10

<210> 200

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 200

tggggggaca

10

<210> 201

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 201

gcggcggggc

10

<210> 202

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 202

gtaggaggat

10

<210> 203

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 203

agggcggggc

10

<210> 204

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 204 gggaaacggg	10
<210> 205 <211> 10 <212> DNA <213> 人工序列	
<220> <223> 合成产生的寡核苷酸	
<400> 205 gtaggagagt	10
<210> 206 <211> 10 <212> DNA <213> 人工序列	
<220> <223> 合成产生的寡核苷酸	
<400> 206 ggaagggtt	10
<210> 207 <211> 10 <212> DNA <213> 人工序列	
<220> <223> 合成产生的寡核苷酸	
<400> 207 gagcaagtag	10
<210> 208 <211> 10 <212> DNA <213> 人工序列	
<220> <223> 合成产生的寡核苷酸	

<400> 208
gaggtgggag 10

<210> 209
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 209
gaggacaatg 10

<210> 210
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 210
ggggcgggat 10

<210> 211
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 211
gagggagaag 10

<210> 212
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 212

gaagagaggt 10

<210> 213

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 213

gagtgagacc

10

<210> 214

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 214

gaggcggaaa

10

<210> 215

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 215

gggaaagaac

10

<210> 216

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 216

gcggaagttc

10

<210> 217

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 217

gagtgggaaa

10

<210> 218

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 218

ggaggggggc

10

<210> 219

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 219

tgggaggatc

10

<210> 220

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 220

gtggggaaaa

10

<210> 221

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 221

gaggttgagg

10

<210> 222

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 222

agagaaggag

10

<210> 223

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 223

tgagatgagc

10

<210> 224

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 224

ggagaggctc

10

<210> 225

<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 225
agggctgatt 10

<210> 226
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 226
ggggaggaaa 10

<210> 227
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 227
agaggaaggt 10

<210> 228
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 228
ggagaagtag 10

<210> 229
<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 229

ggtggcaggt

10

<210> 230

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 230

gctggagggg

10

<210> 231

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 231

gcggggaggg

10

<210> 232

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 232

aaagtgggga

10

<210> 233

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 233

agaaaaaata

10

<210> 234

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 234

gacaggggag

10

<210> 235

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 235

tgagttggga

10

<210> 236

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 236

gaaggaaaat

10

<210> 237

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 237

ggggctgagg

10

<210> 238

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 238

tgagacggag

10

<210> 239

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 239

gctggaaatt

10

<210> 240

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 240

caagtagaag

10

<210> 241

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 241

gaggcggaaa

10

<210> 242

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 242

gctgcagcgt

10

<210> 243

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 243

gatggggttt

10

<210> 244

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 244

gaagcggagg

10

<210> 245

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 245

gtggcgggaag

10

<210> 246

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 246

aaaggaaaag

10

<210> 247

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 247

cgggttaaaa

10

<210> 248

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 248

gtagctggga

10

<210> 249

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 249

ggacaagcgg

10

<210> 250

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 250

aaaagaaaaa

10

<210> 251

<211> 131

<212> DNA

<213> 人工序列

<220>

<223> 载体序列

<400> 251

gaccaagct tgccaccatg gtgtaccct acgacgtgcc cgactacgcc gaattgcctc
caaaaaagaa gagaaaggta gggatccgaa ttcaagcggc cgcagatgat ctcgagcatg
catctagagg g

60

120

131

<210> 252

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 252

agagtagaat

10

<210> 253

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 253

aaagtaaaaa

10

<210> 254

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 254

agggagggag

10

<210> 255

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 255

aaaaatgagc

10

<210> 256

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 256

cgggaaagaa

10

<210> 257

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 257
gtagcaagac 10

<210> 258
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 258
aatgtaaaaa 10

<210> 259
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 259
cggggaggagg 10

<210> 260
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 260
caaagagact 10

<210> 261
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 261
gaagatgcgt 10

<210> 262
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 262
cgagcatggg 10

<210> 263
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 263
acaacagggg 10

<210> 264
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 264
gttggggggc 10

<210> 265
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 265

agggaggtgt	10
<210> 266	
<211> 10	
<212> DNA	
<213> 人工序列	
<220>	
<223> 合成产生的寡核苷酸	
<400> 266	
tgagacgggg	10
<210> 267	
<211> 10	
<212> DNA	
<213> 人工序列	
<220>	
<223> 合成产生的寡核苷酸	
<400> 267	
gaagttggaa	10
<210> 268	
<211> 10	
<212> DNA	
<213> 人工序列	
<220>	
<223> 合成产生的寡核苷酸	
<400> 268	
agaaaagaaa	10
<210> 269	
<211> 10	
<212> DNA	
<213> 人工序列	
<220>	
<223> 合成产生的寡核苷酸	
<400> 269	
gactgacaat	10

<210> 270

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 270

gctgaggatg

10

<210> 271

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 271

ggggagaaat

10

<210> 272

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 272

tgatgagaaa

10

<210> 273

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 273

gcaggagact

10

<210> 274

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 274

tggggagatt

10

<210> 275

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 275

gcggaatgga

10

<210> 276

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 276

gtctggggac

10

<210> 277

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 277

gagggggcgt

10

<210> 278

<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 278
gacgctgctt 10

<210> 279
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 279
ggggcaggcg 10

<210> 280
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 280
aaaaaaaaag 10

<210> 281
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 合成产生的寡核苷酸

<400> 281
ggaagagagg 10

<210> 282
<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 282

caagggaaaa

10

<210> 283

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 283

ggtgatgcac

10

<210> 284

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 284

aaggtcgcgg

10

<210> 285

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 285

ggggctggag

10

<210> 286

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 286

gggggtgtac

10

<210> 287

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 287

ggtgatgctc

10

<210> 288

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 288

gtttgagaca

10

<210> 289

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 289

gctaaaaatc

10

<210> 290

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 290

agaggagctt

10

<210> 291

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 291

tgagatgggg

10

<210> 292

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 292

gcttggggct

10

<210> 293

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 293

gttggggcgg

10

<210> 294

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 294

ggagctgctt

10

<210> 295

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 295

gatgcaggac

10

<210> 296

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 296

ggccgggtcg

10

<210> 297

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 297

gatggtggtt

10

<210> 298

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 298

gccggggcgg

10

<210> 299

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 299

gctggggact

10

<210> 300

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 300

gtagctgtaa

10

<210> 301

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 301

gggggcggtt

10

<210> 302

<211> 10

<212> DNA

<213> 人工序列

<220>

<223> 合成产生的寡核苷酸

<400> 302 ggtgctgatt	10
<210> 303 <211> 10 <212> DNA <213> 人工序列	
<220> <223> 合成产生的寡核苷酸	
<400> 303 gcagtaggag	10
<210> 304 <211> 10 <212> DNA <213> 人工序列	
<220> <223> 合成产生的寡核苷酸	
<400> 304 gacgaagggg	10
<210> 305 <211> 10 <212> DNA <213> 人工序列	
<220> <223> 合成产生的寡核苷酸	
<400> 305 gacgacgctt	10

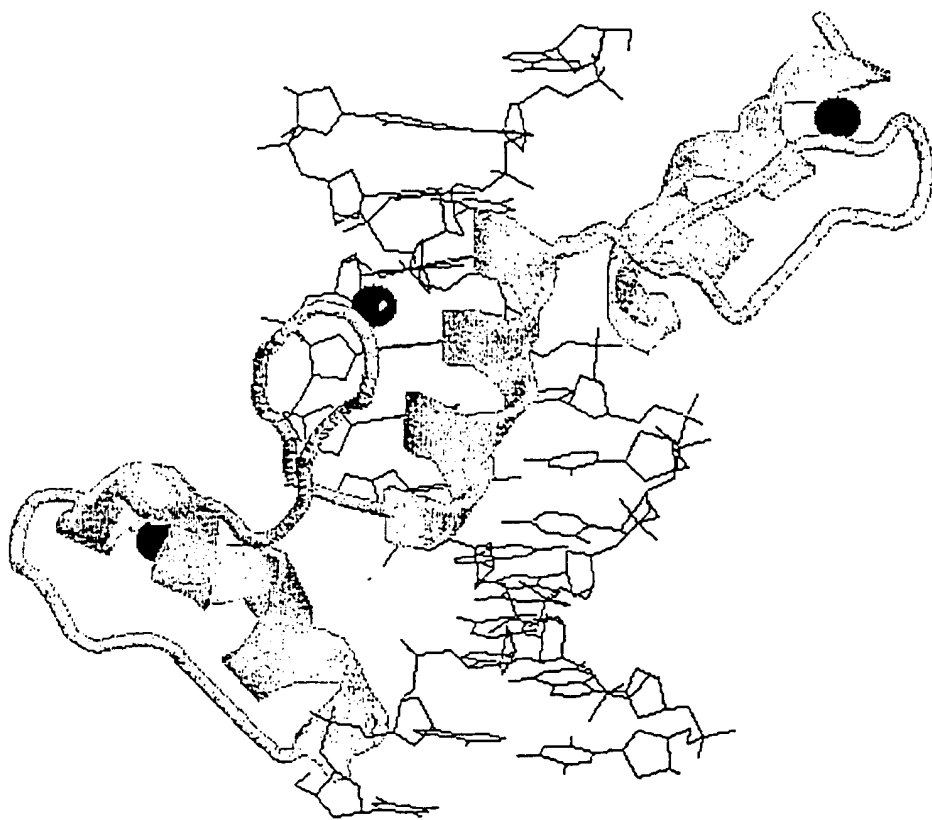
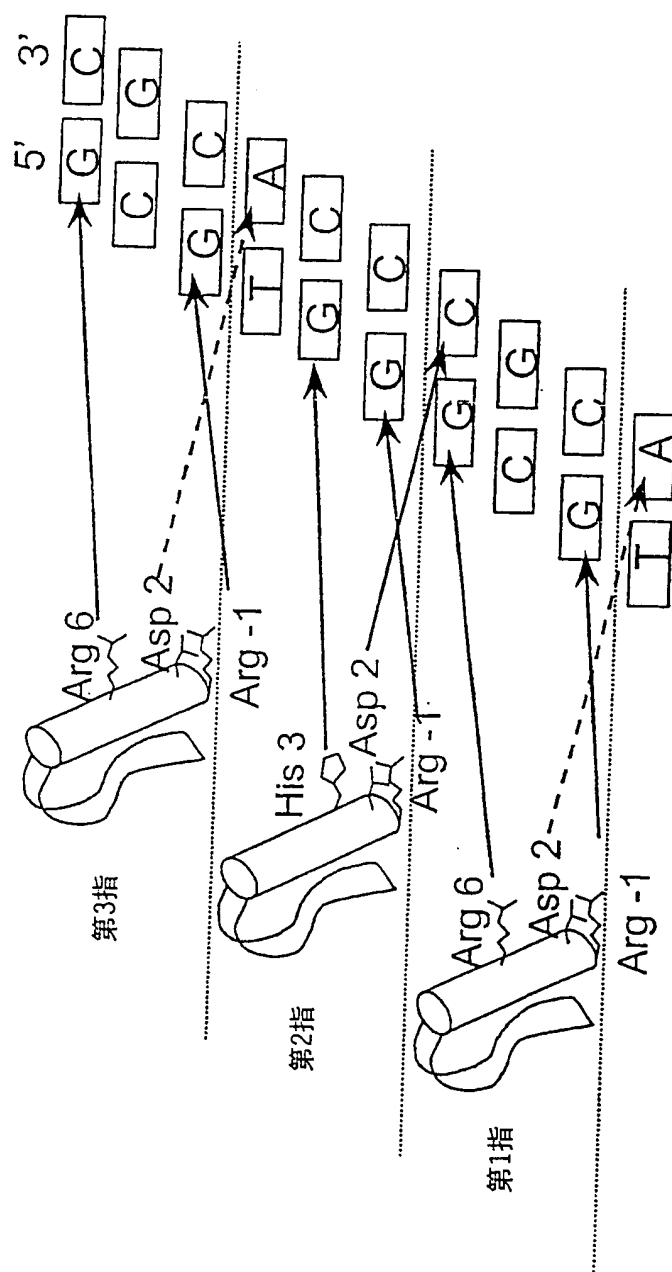


图1



2
圖

碱基

	G	A	C	T
5'	Arg6 Lys6 Asp2 Ser2	Gln6	Ser2	Lys6 Asp2
在三联体中的位置	His3 Lys3	Asn3 Ser3 His3	Asp3 Thr3 Val3 Leu3	Thr3 Ala3 Ser3 Val3
3'	Arg-1	Gln-1	Asp-1	Leu-1 Thr-1 Asn-1

图3

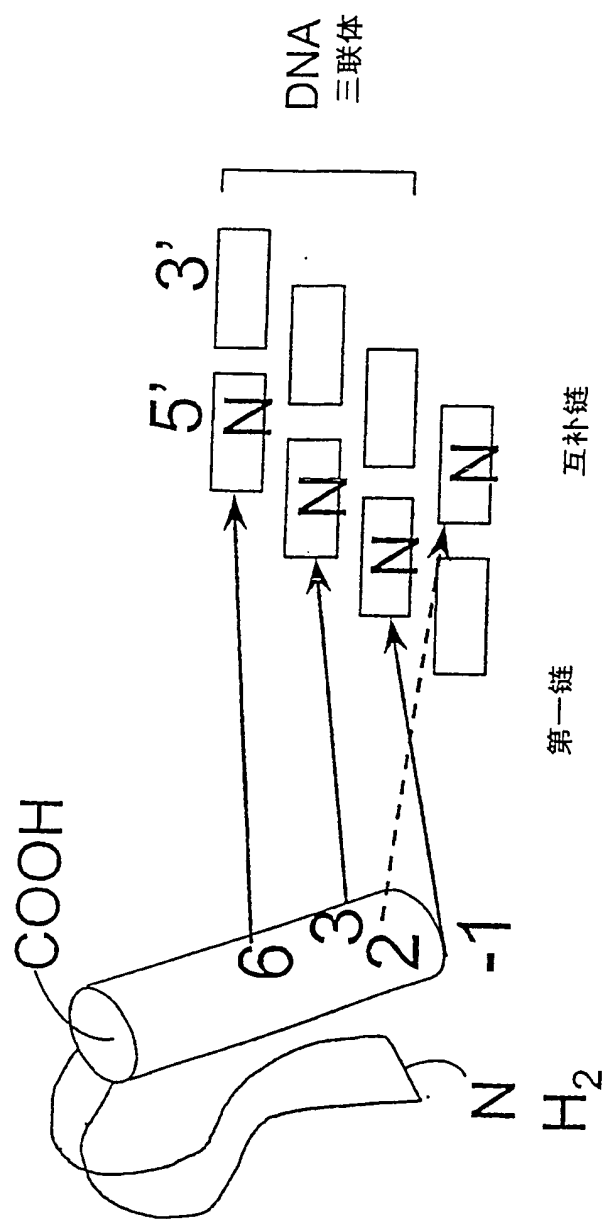
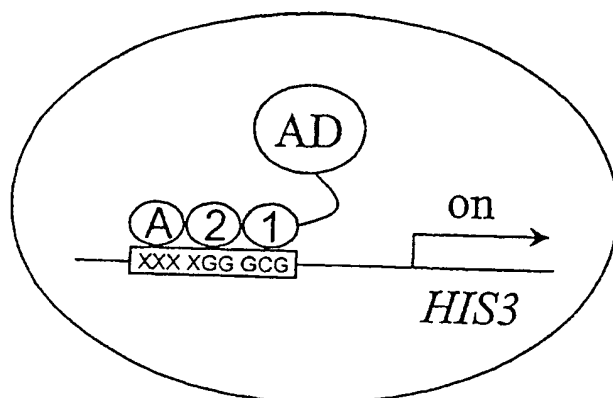
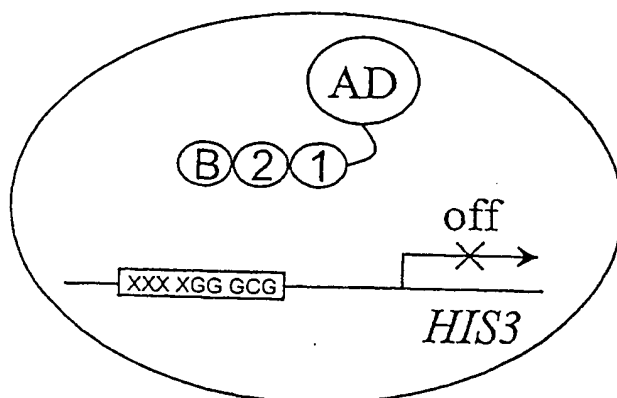


图4



是

在无组氨酸平板上的生长



否

在无组氨酸平板上的生长

图5

HIV-1 LTR (-124/-115): 5'-GAC ATC GAG C-3' (SEQ ID NO:1)

HIV-1 LTR (-23/-14): 5'-GCA GCT GCT T-3' (SEQ ID NO:2)

HIV-1 LTR (-95/-86): 5'-GCT GGG GAC T-3' (SEQ ID NO:3)

人 CCR5 (-70/-79): 5'-AGG GTG GAG T-3' (SEQ ID NO:4)

人 CCR5 (+7/+16): 5'-GCT GAG ACA T-3' (SEQ ID NO:5)

图6

GCGT (最佳 Zif268- 结合位点)

5'-CCG GCGTGGGCG GCT GCGTGGGCG T GCGTGGGCG GACT GCGTGGGCG-3' (SEQ ID NO:6)
3'-GCACCCGC CGA CGACCCGC A CGACCCGC CTGA CGACCCGC AGCT-5' (SEQ ID NO:7)

GAGC (HIV-1 LTR, -118/-115)

5'-CCGGC GAGCGGGCG GTC GAGCGGGCG T GAGCGGGCG GATC GAGCGGGCG-3' (SEQ ID NO:8)
3'-G CTCGCCCCG CAG CTCGCCCCG A CTCGCCCCG CTAG CTCGCCCCG AGCT-5' (SEQ ID NO:9)

GCTT (HIV-1 LTR, -17/-14)

5'-CCGGCT GCTTGGGCG GCT GCTTGGGCG T GCTTGGGCG GGCT GCTTGGGCG-3' (SEQ ID NO:10)
3'-GA CGAACCCGC CGA CGAACCCGC A CGAACCCGC CCGA CGAACCCGC AGCT-5' (SEQ ID NO:11)

GACT (HIV-1 LTR, -89/-86)

5'-CCG GACTGGGCG GGG GACTGGGCG T GACTGGGCG GAGG GACTGGGCG-3' (SEQ ID NO:12)
3'-TGACCCGC CCC CTGACCCGC A CTGACCCGC CTCC CTGACCCGC AGCT-5' (SEQ ID NO:13)

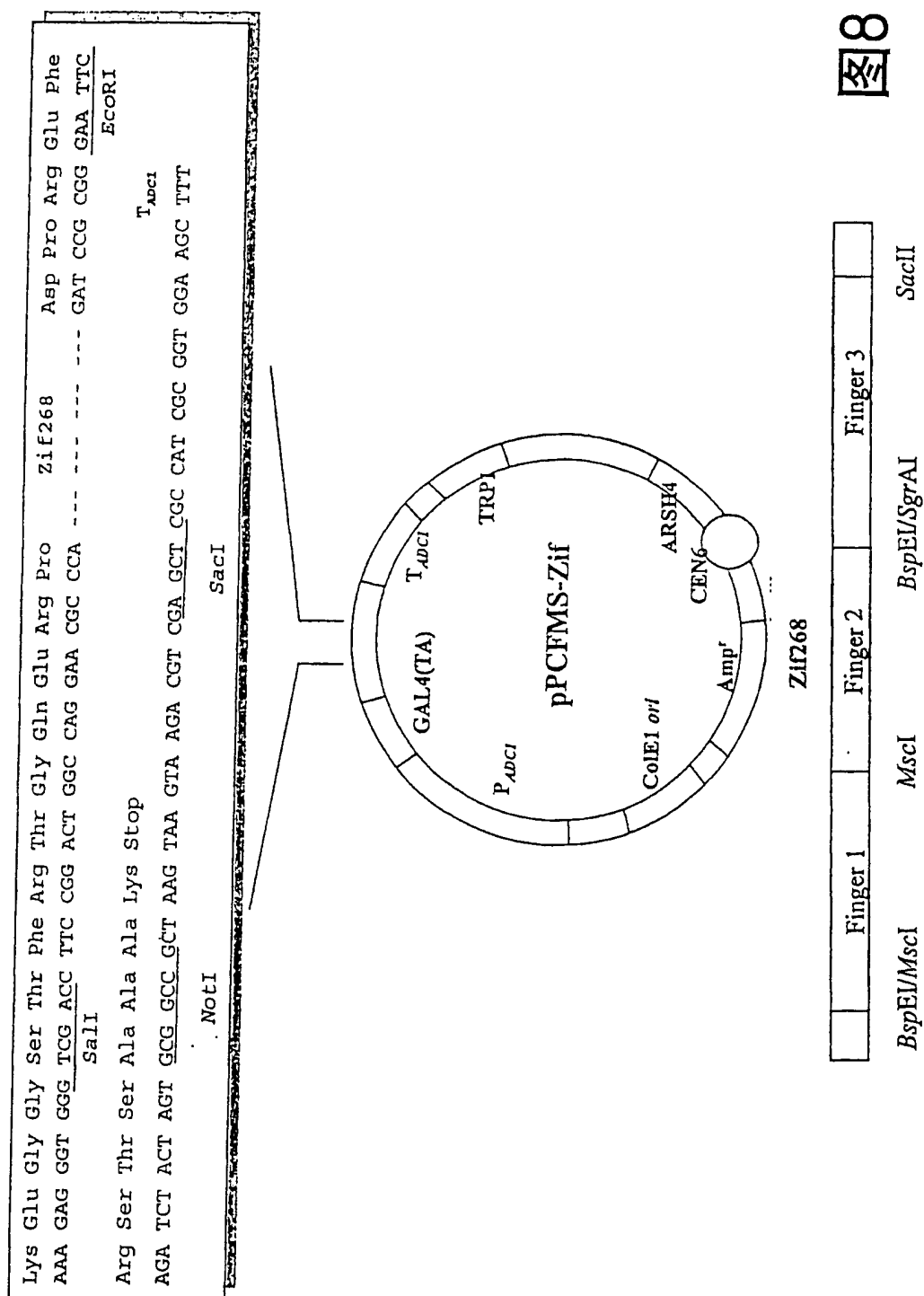
GAGT (人 CCR5, -76/-79)

5'-CCG GAGTGGGCG GTG GAGTGGGCG T GAGTGGGCG GATG GAGTGGGCG-3' (SEQ ID NO:14)
3'-TCACCCGC CAC CTCACCCGC A CTCACCCGC CTAC CTCACCCGC AGCT-5' (SEQ ID NO:15)

ACAT (人 CCR5, +13/+16)

5'-CCGG ACATGGGCG GAG ACATGGGCG T ACATGGGCG GAAG ACATGGGCG-3' (SEQ ID NO:16)
3'-TGTACCCGC CTC TGTACCCGC A TGTACCCGC CTTC TGTACCCGC AGCT-5' (SEQ ID NO:17)

图7



E R P Y A C - P V E S C
 GGG TCG ACC TTC CGG ACT GGC CAG GAA CGC CCA TAT GCT TGC CCT GTC GAG TCC TGC
Sall *BspEI* *MscI*

D R R F S R S D E L T R H I R I H T G
 GAT CGC CGC TTT TCT CGC TCG GAT GAG CTT ACC CGC CAT ATC CGC ATC CAC ACT GGC
MscI

Q K P F Q C R I C M R N F S R S D H L
 CAG AAG CCC TTC CAG TGT CGA ATC TGC ATG CGT AAC TTC AGT CGT AGT GAC CAC CTT

T T H I R T H T G E K P F A C D I C G
 ACC ACC CAC ATC CGG ACC CAC ACC GGC GAG AAG CCT TTT GCC TGT GAC ATT TGT GGG
BspEI *SgrAI*

R K F A R S D E R K R H T K I H L R Q
 AGG AAG TTT GCC AGG AGT GAT GAA CGC AAG AGG CAT ACC AAA ATC CAT TTA AGA CAG

K D (SEQ ID NO: 21)
 AAG GAT CCG CGG GAA TCC (SEQ ID NO: 20)
SacII *EcoRI*

图9

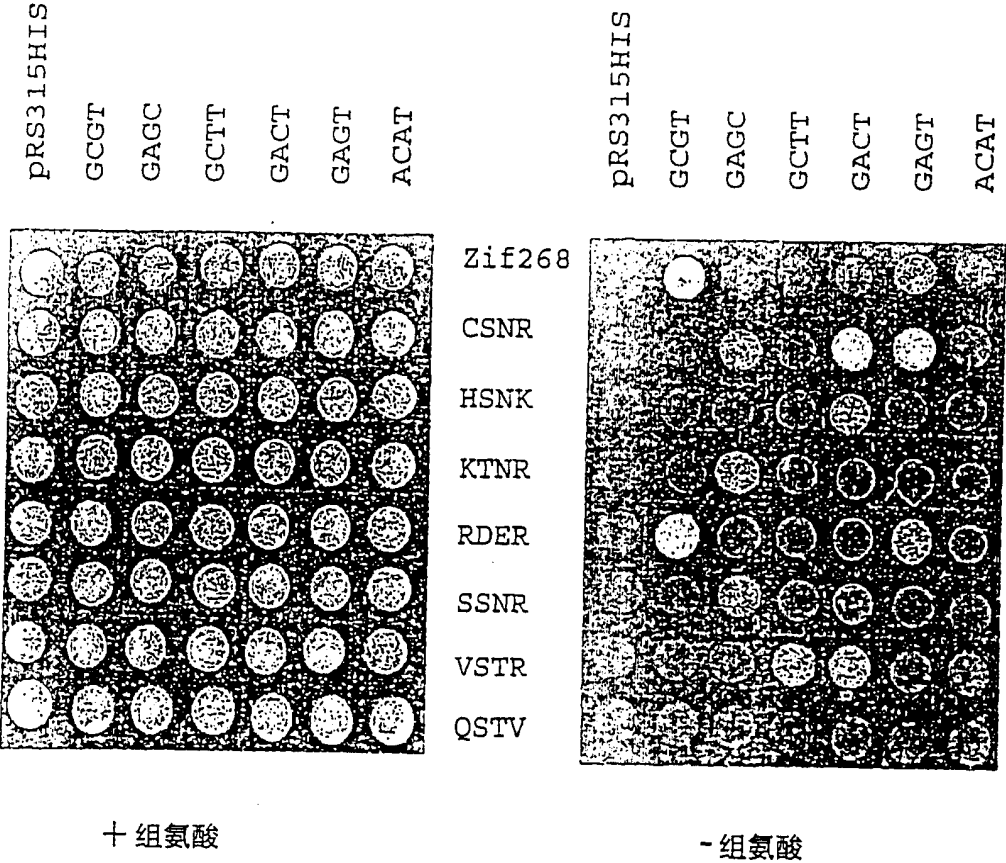


图10

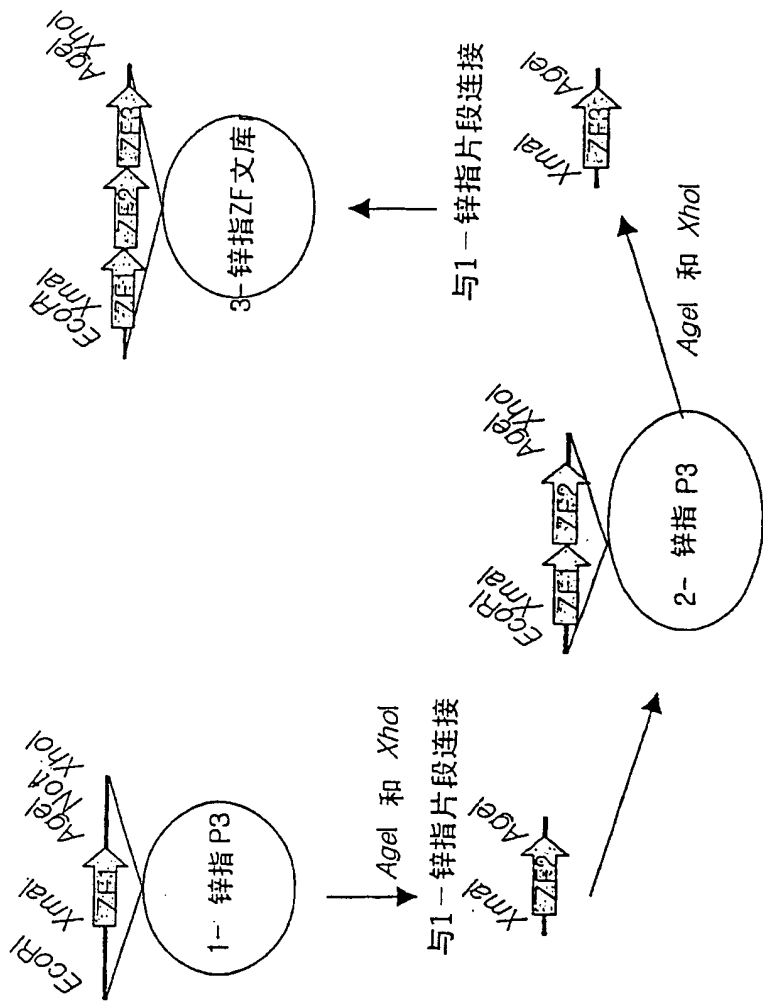


图11B

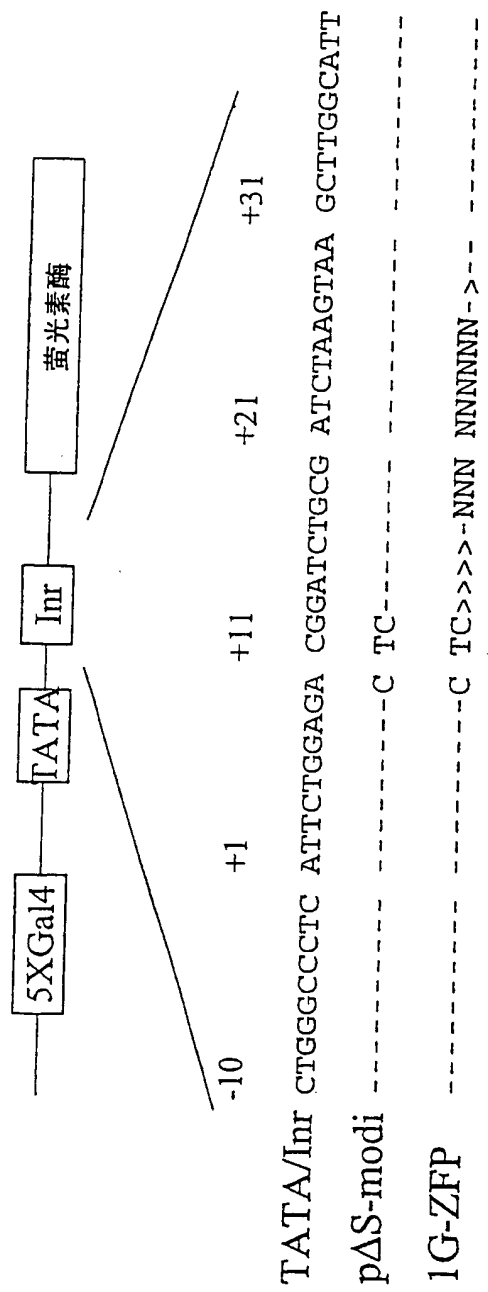


图12