US 20170356037A9

(19) **United States**

(12) **Patent Application Publication**

Clarke et al.

(10) **Pub. No.: US 2017/0356037 A9**

(48) **Pub. Date: Dec. 14, 2017**

CORRECTED PUBLICATION

(54) **MUTANT PORES**

(71) Applicant: **Oxford Nanopore Technologies Ltd.**, Oxford (GB)

(72) Inventors: **James Anthony Clarke**, Oxford (GB); **Andrew John Heron**, Oxford (GB); **Lakmal Jayasinghe**, Oxford (GB); **Elizabeth Jayne Wallace**, Oxford (GB)

(73) Assignee: **Oxford Nanopore Technologies Ltd.**, Oxford (GB)

(21) Appl. No.: **15/308,206**

(22) PCT Filed: **May 1, 2015**

(86) PCT No.: **PCT/GB2015/051290**

§ 371 (c)(1),
(2) Date: **Nov. 1, 2016**

Prior Publication Data

(15) Correction of US 2017/0058337 A1 Mar. 2, 2017
See Claims 4 and 5.

(65) US 2017/0058337 A1 Mar. 2, 2017

(30) **Foreign Application Priority Data**

May 2, 2014 (GB) 1407809.1
Oct. 7, 2014 (GB) 1417708.3
Oct. 7, 2014 (GB) 1417712.5

Publication Classification

(51) **Int. Cl.**
C12Q 1/68 (2006.01)
C07K 14/35 (2006.01)

(52) **U.S. Cl.**
CPC ***C12Q 1/6869*** (2013.01); ***C07K 14/35*** (2013.01)

(57) **ABSTRACT**

The invention relates to mutant forms of Msp. The invention also relates to polynucleotide characterisation using Msp.

Figure 1

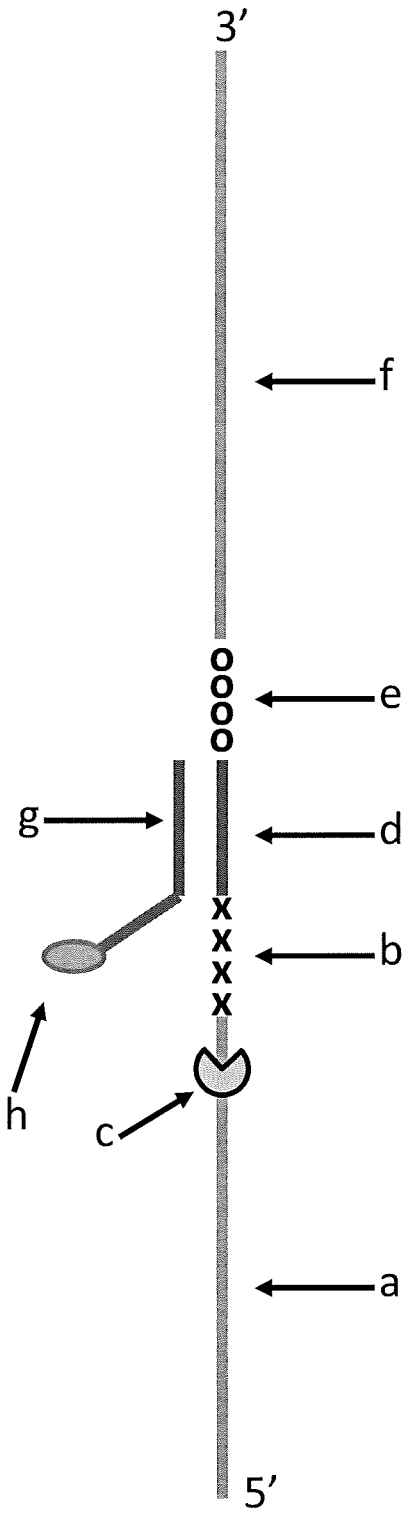


Figure 2

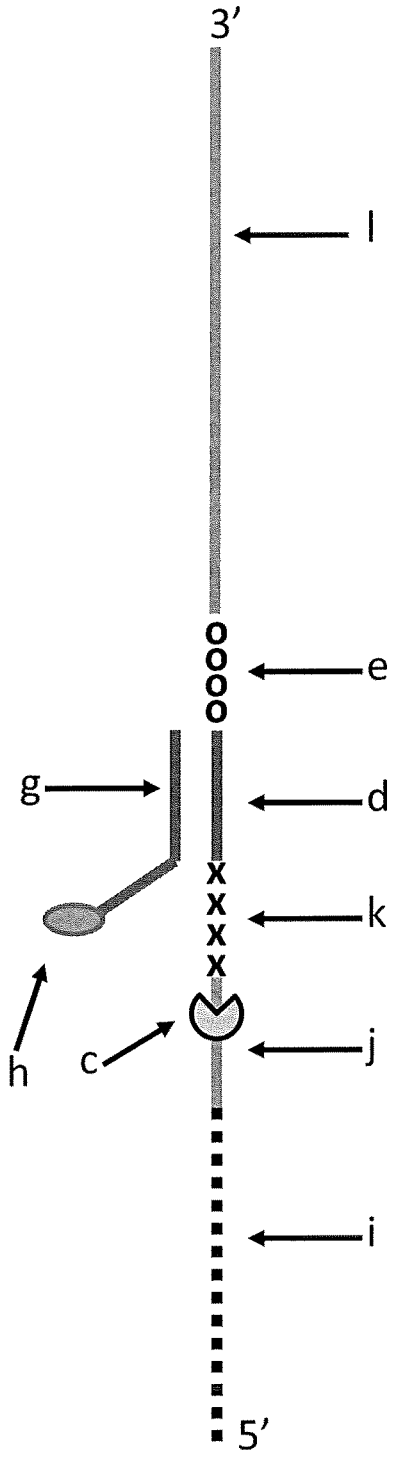


Figure 3

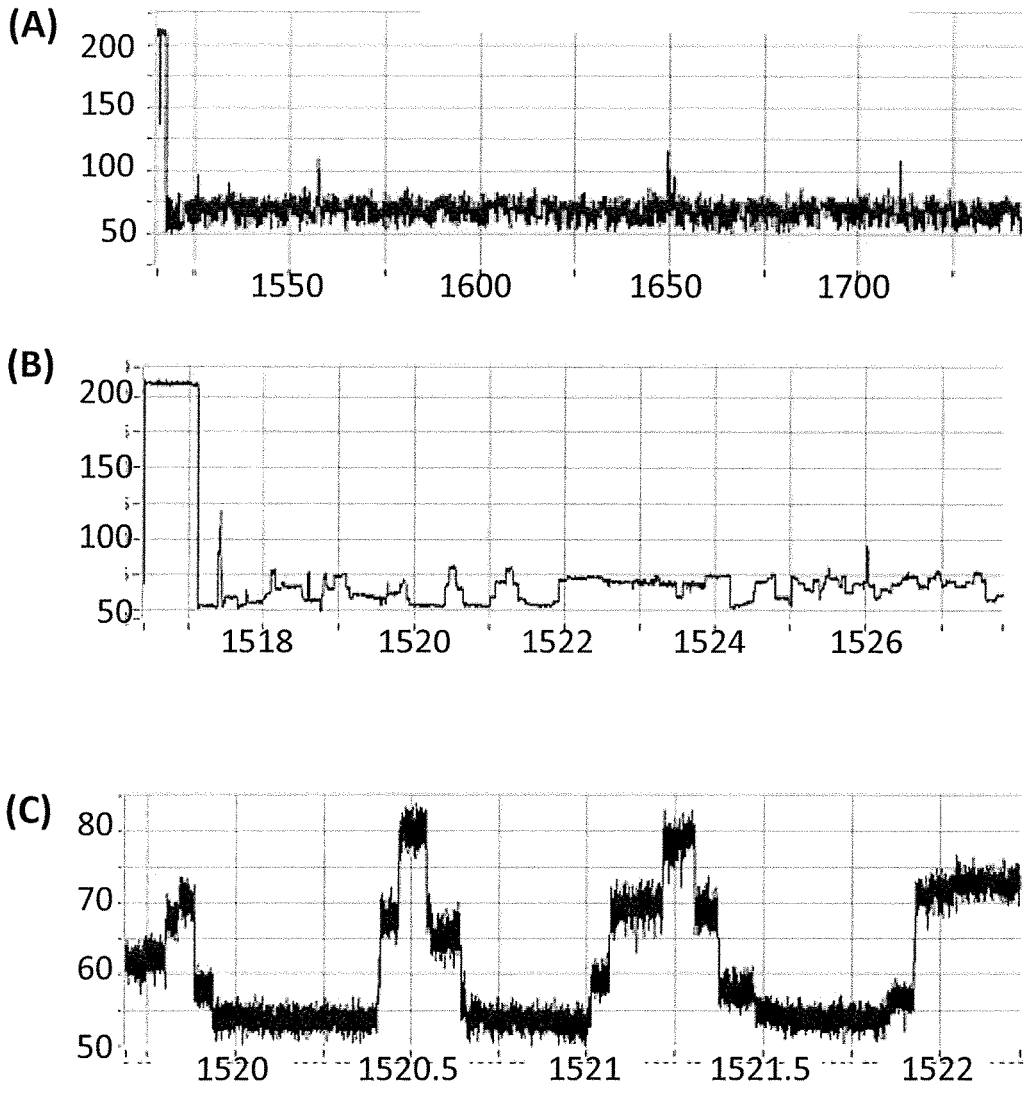

(A)
200
150
100
50
1550
1600
1650
1700
(B)
200
150
100
50
1518
1520
1522
1524
1526
(C)
80
70
60
50
1520
1520.5
1521
1521.5
1522

Figure 4

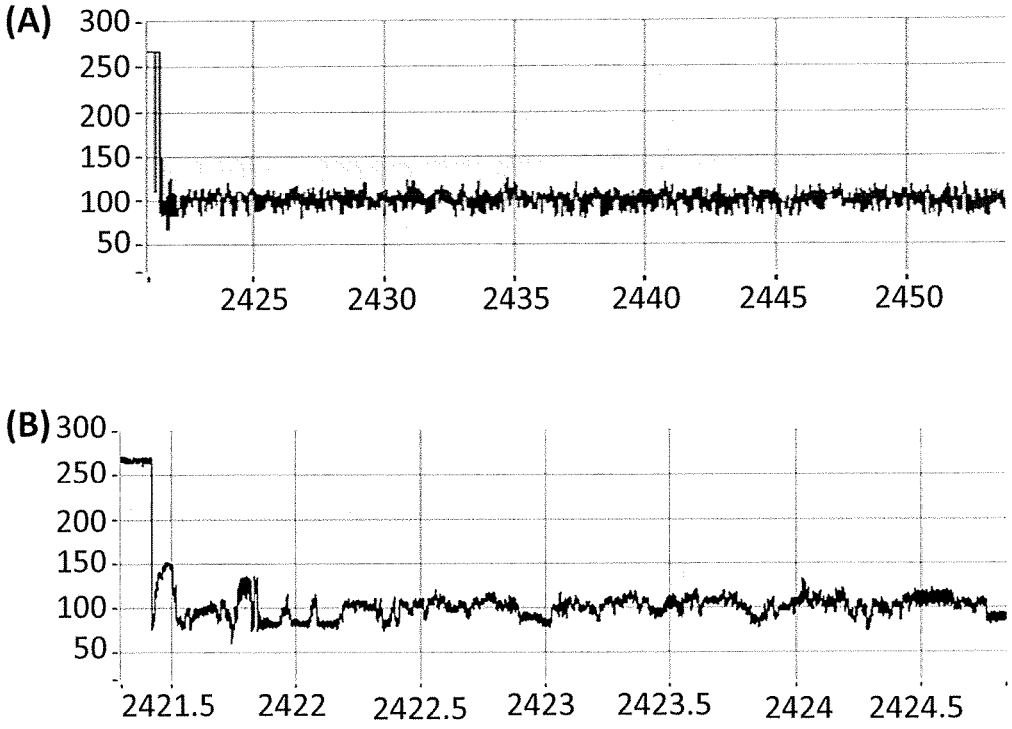

(A)
300
250
200
150
100
50
2425
2430
2435
2440
2445
2450
(B)
300
250
200
150
100
50
2421.5
2422
2422.5
2423
2423.5
2424
2424.5

Figure 5

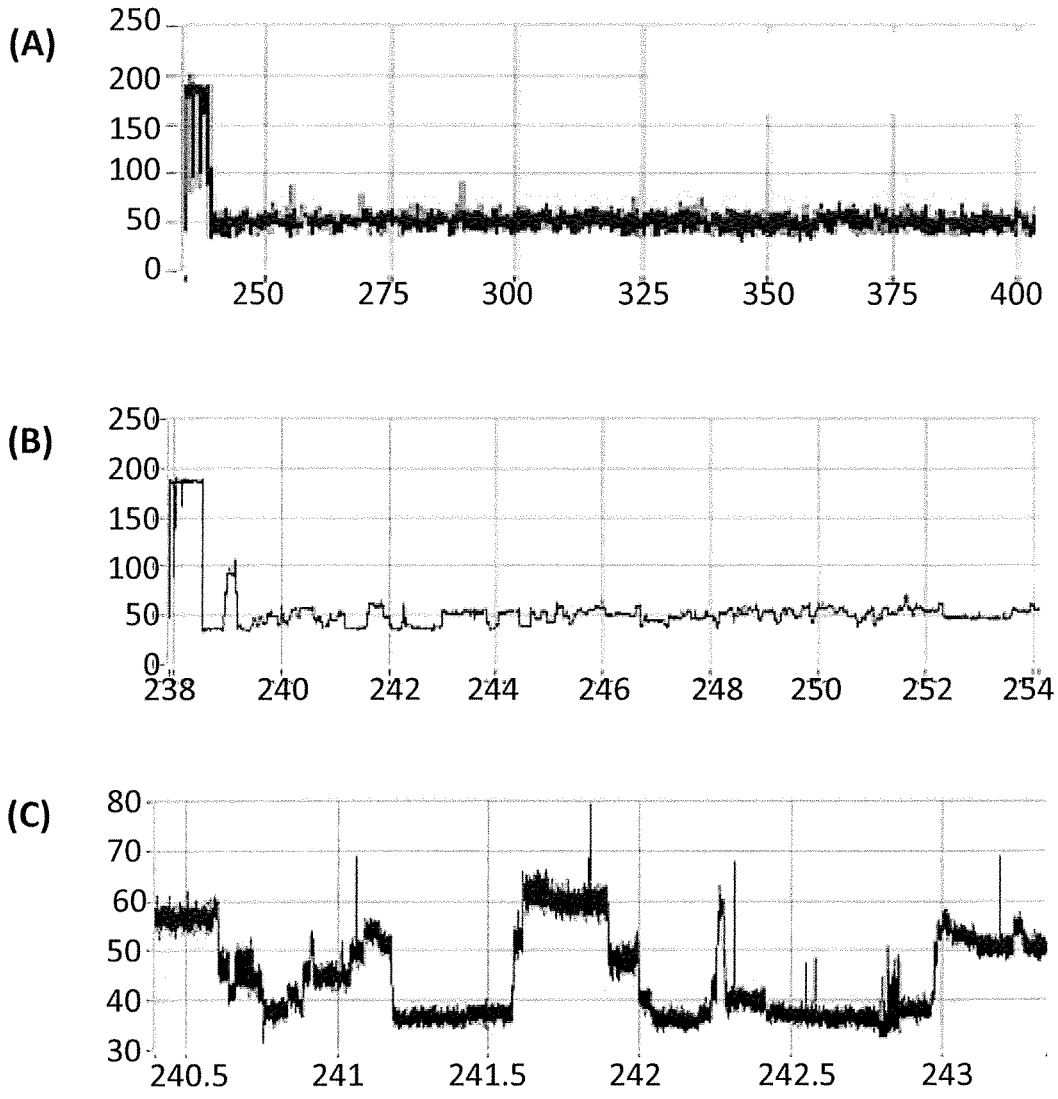

(A)
250
200
150
100
50
0
250
275
300
325
350
375
400
(B)
250
200
150
100
50
0
238
240
242
244
246
248
250
252
254
(C)
80
70
60
50
40
30
240.5
241
241.5
242
242.5
243

Figure 6

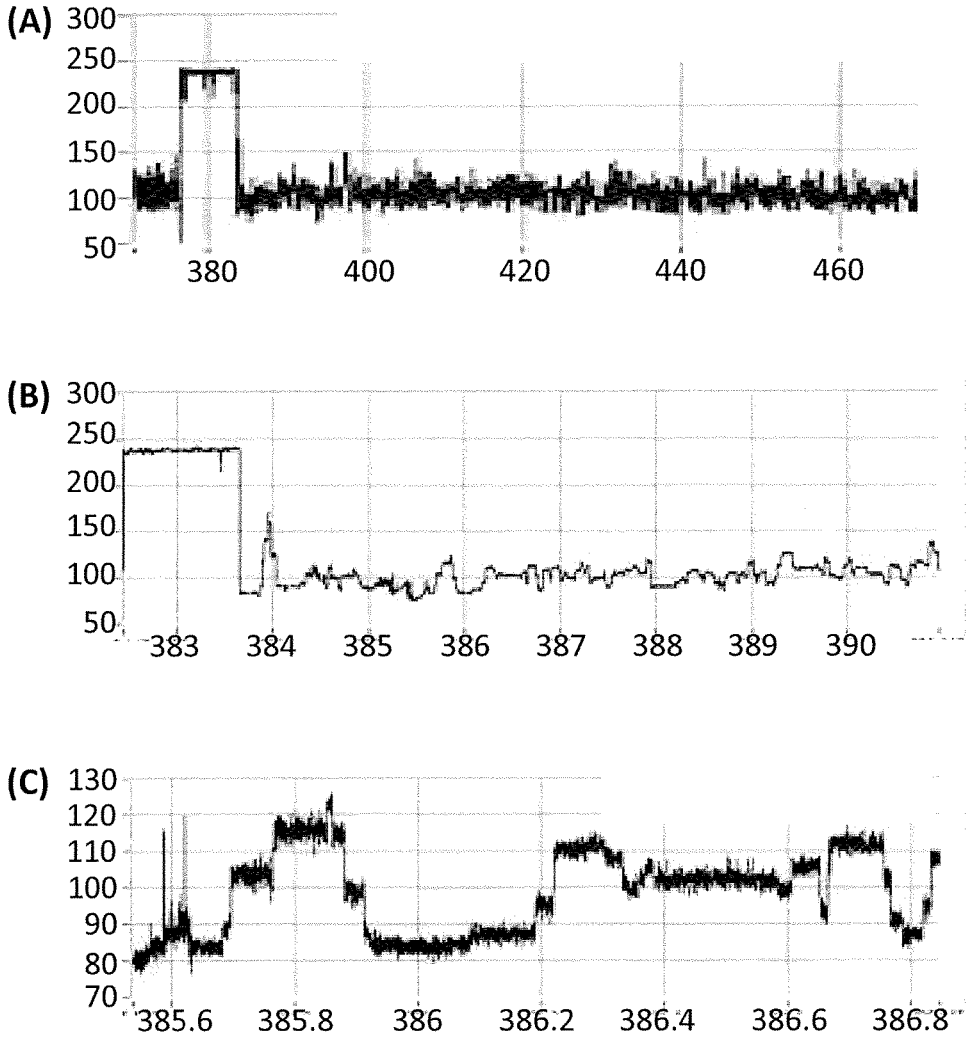

(A)
300
250
200
150
100
50
380
400
420
440
460
(B)
300
250
200
150
100
50
383
384
385
386
387
388
389
390
(C)
130
120
110
100
90
80
70
385.6
385.8
386
386.2
386.4
386.6
386.8

Figure 7

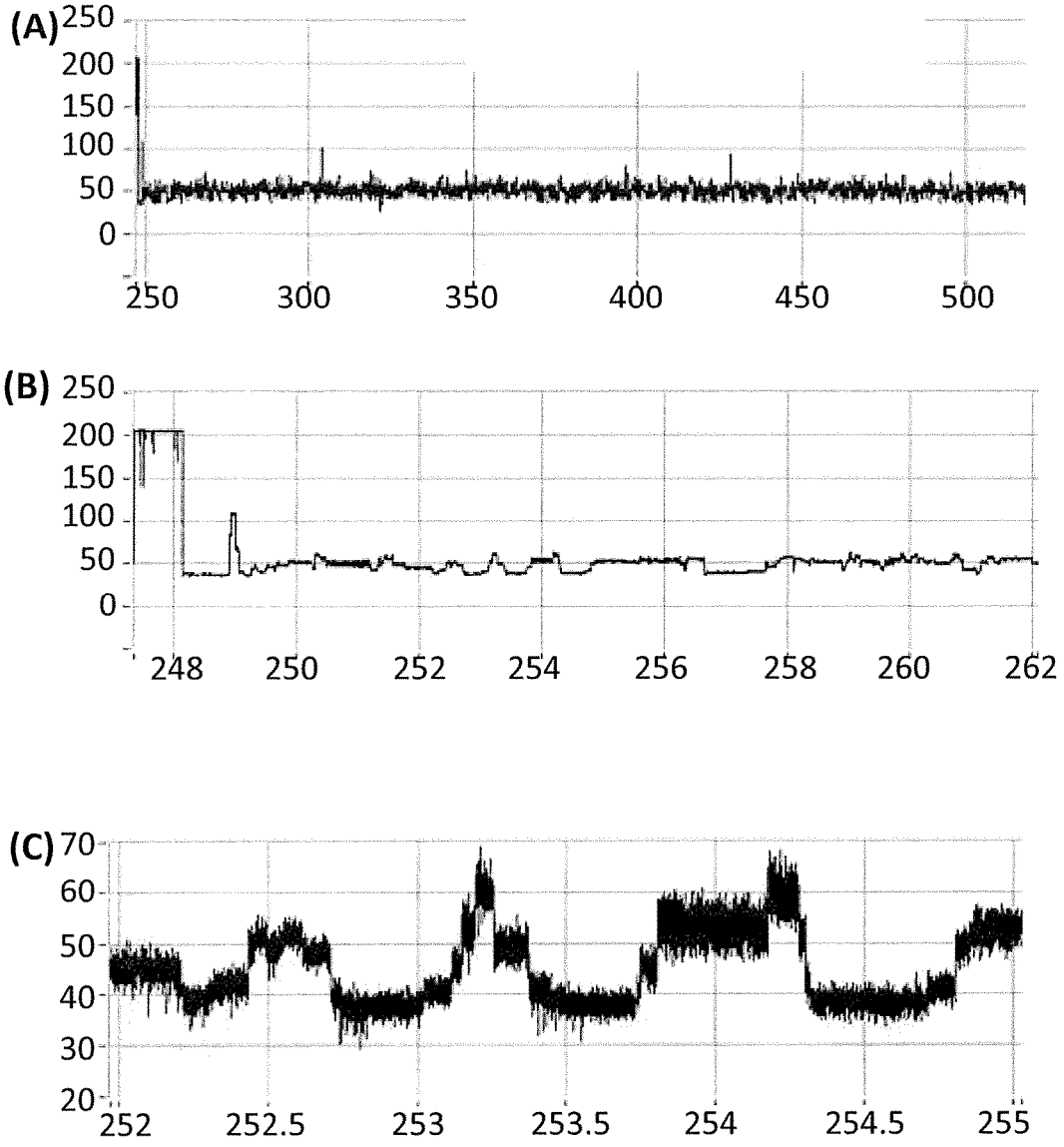

(A)
250
200
150
100
50
0
250
300
350
400
450
500
(B)
250
200
150
100
50
0
248
250
252
254
256
258
260
262
(C)
70
60
50
40
30
20
252
252.5
253
253.5
254
254.5
255

Figure 8

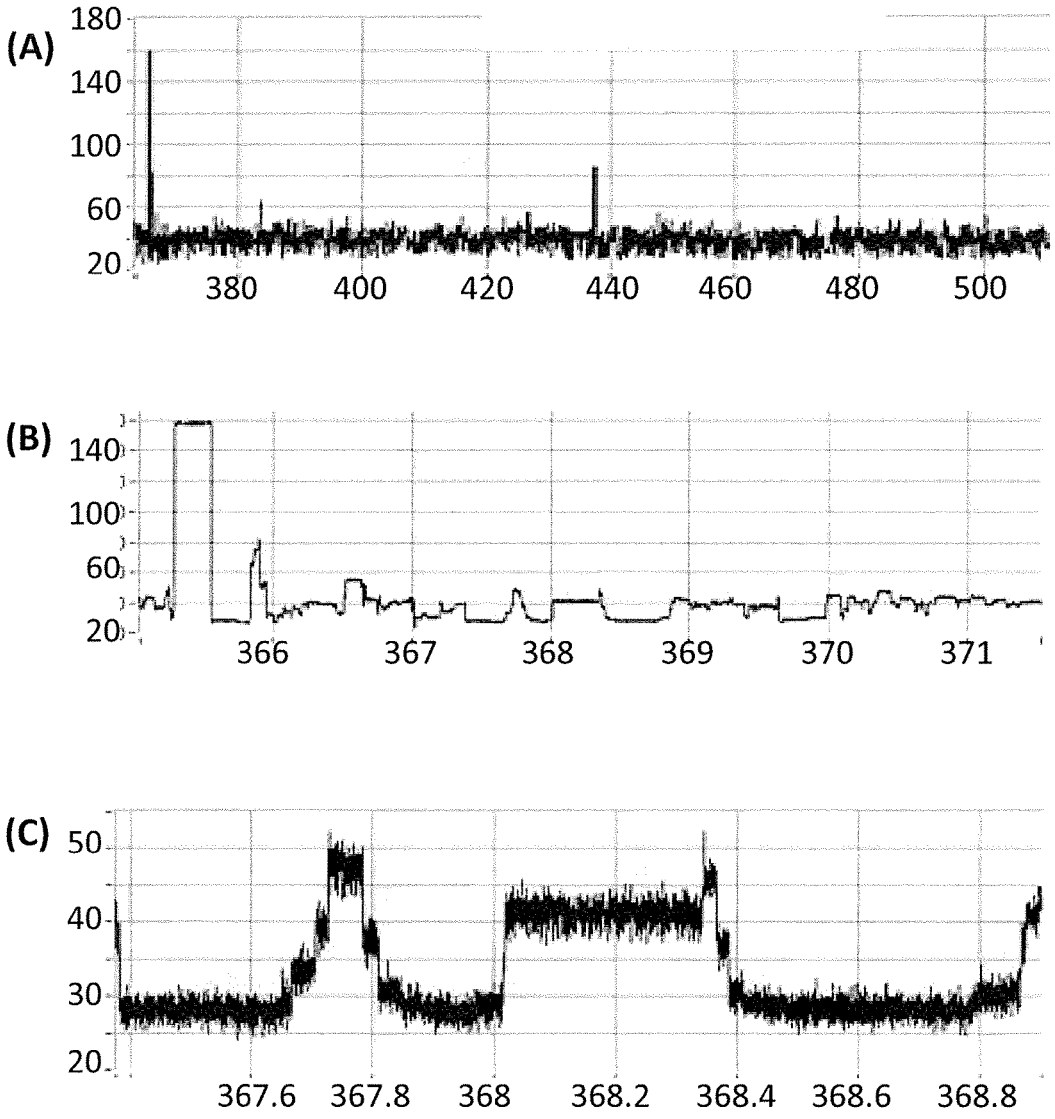

(A)
180
140
100
60
20
380
400
420
440
460
480
500
(B)
140
100
60
20
366
367
368
369
370
371
(C)
50
40
30
20
367.6
367.8
368
368.2
368.4
368.6
368.8

Figure 9

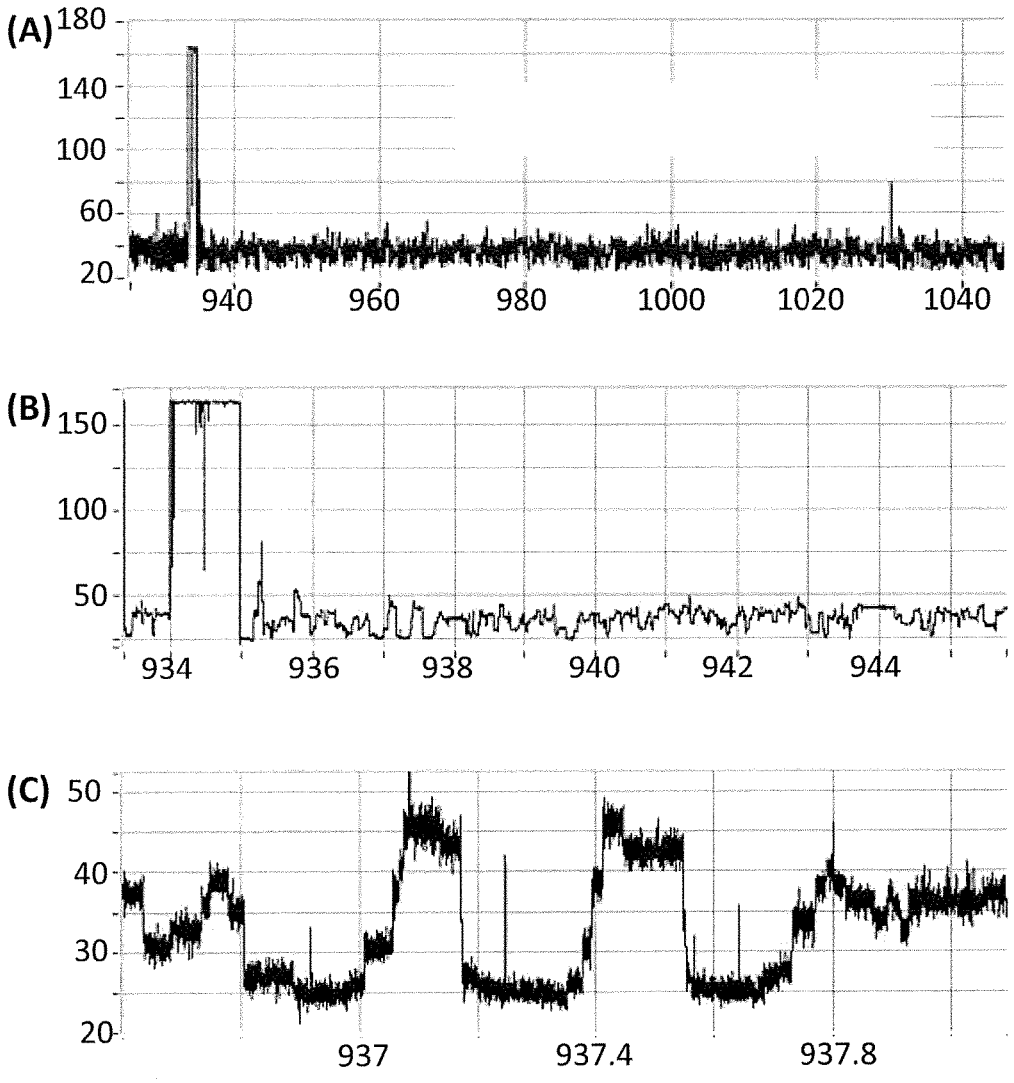

(A)
180
140
100
60
20
940
960
980
1000
1020
1040
(B)
150
100
50
934
936
938
940
942
944
(C)
50
40
30
20
937
937.4
937.8

Figure 10

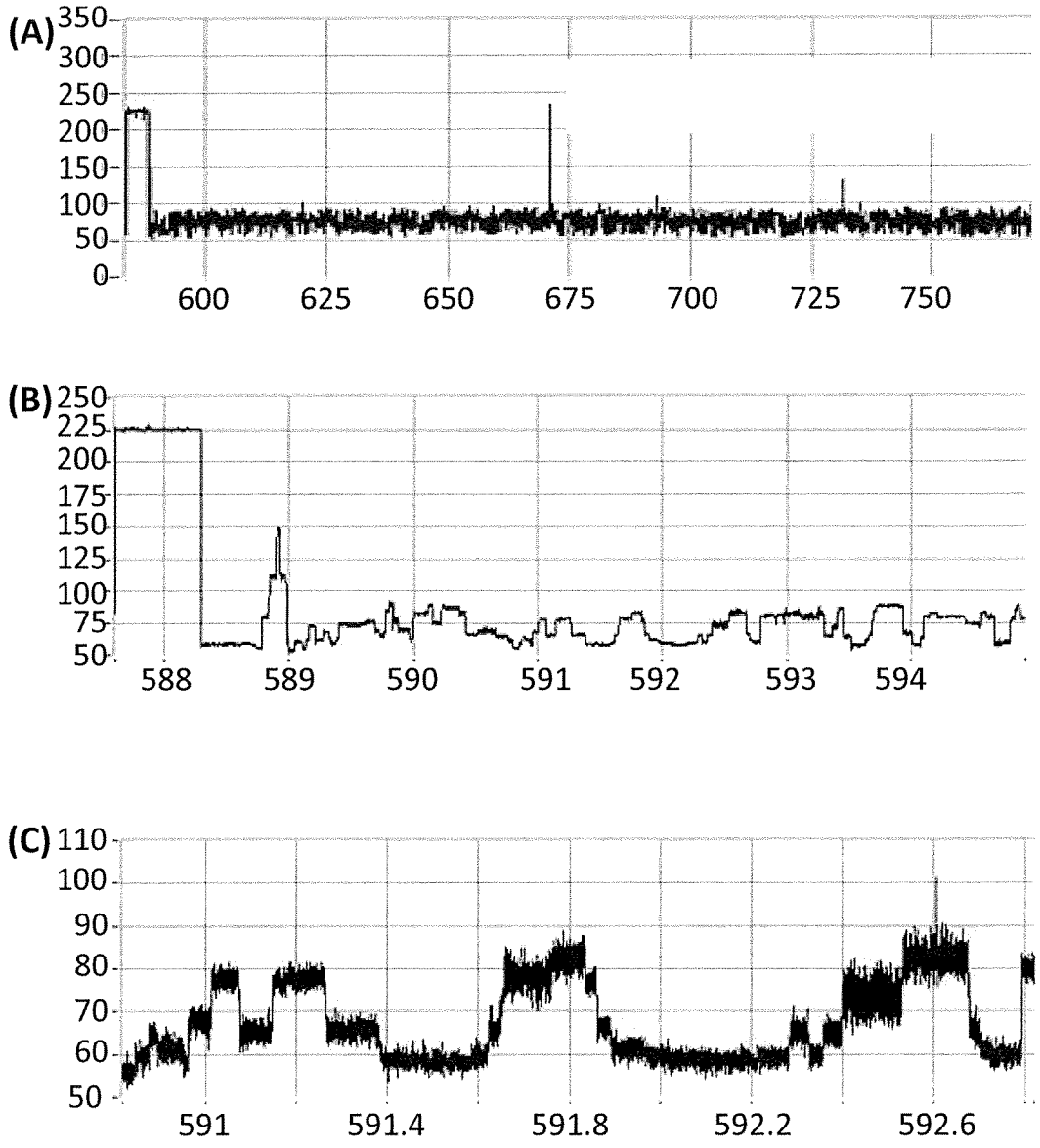

(A)
350
300
250
200
150
100
50
0
600
625
650
675
700
725
750
(B)
250
225
200
175
150
125
100
75
50
588
589
590
591
592
593
594
(C)
110
100
90
80
70
60
50
591
591.4
591.8
592.2
592.6

Figure 11

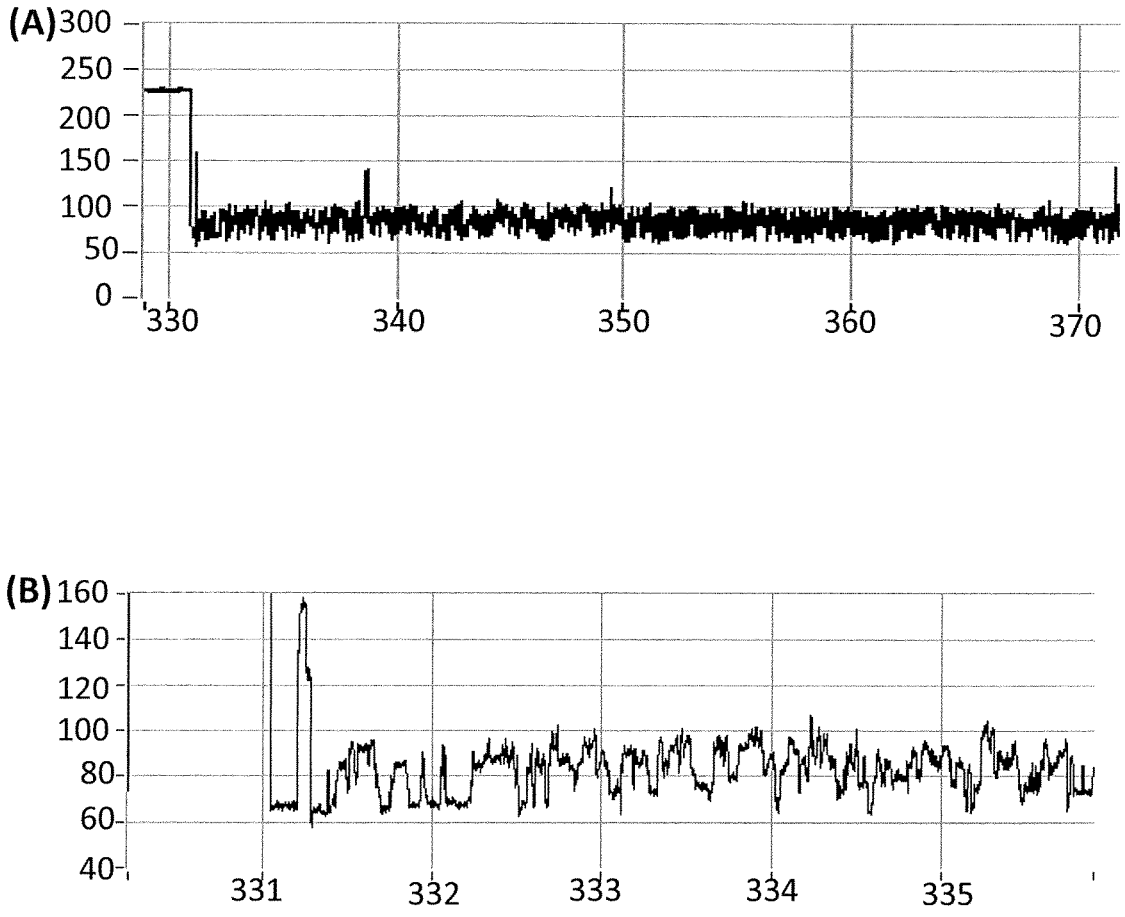

(A)
300
250
200
150
100
50
0
330
340
350
360
370
(B)
160
140
120
100
80
60
40
331
332
333
334
335

Figure 12

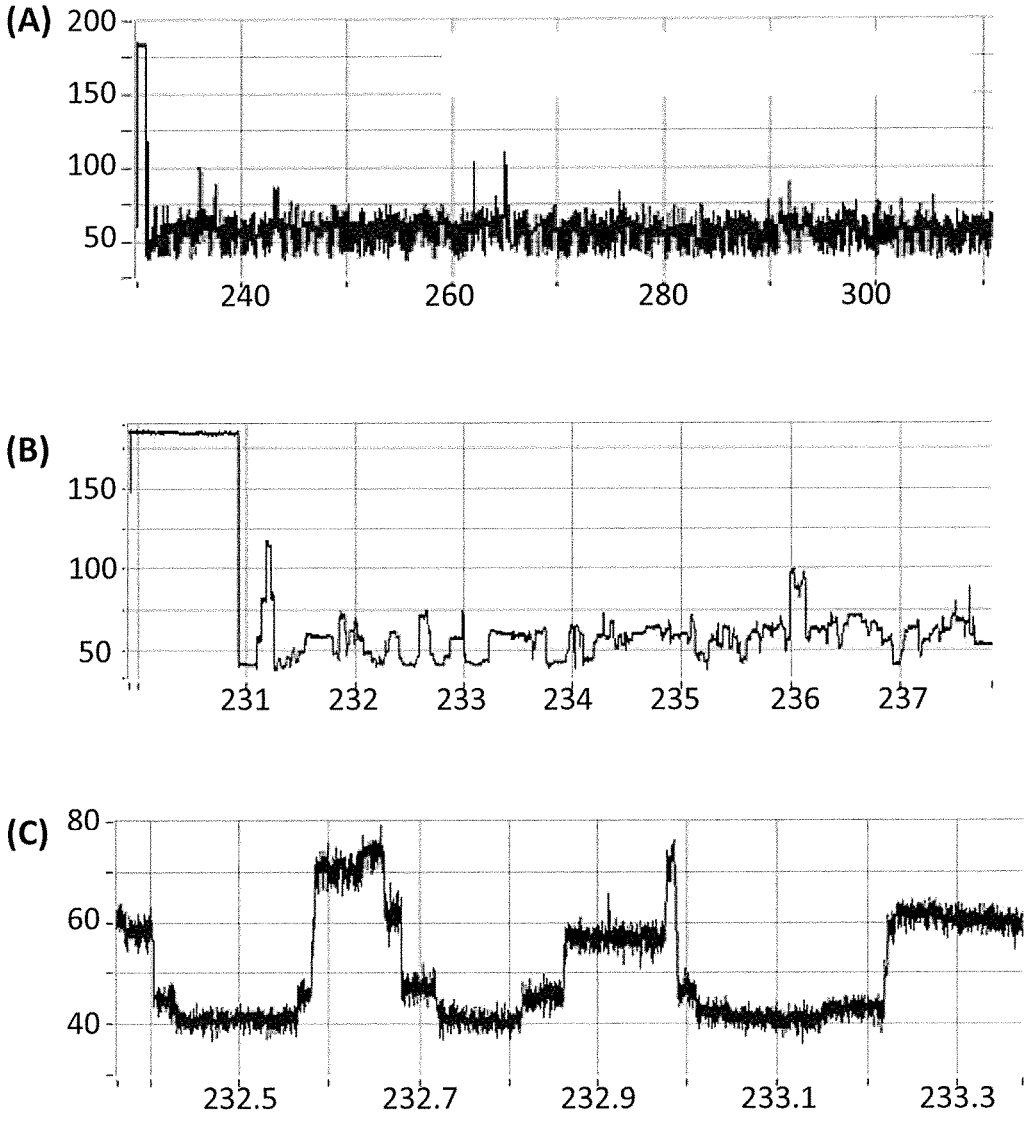

(A)
200
150
100
50
240
260
280
300
(B)
150
100
50
231
232
233
234
235
236
237
(C)
80
60
40
232.5
232.7
232.9
233.1
233.3

Figure 13

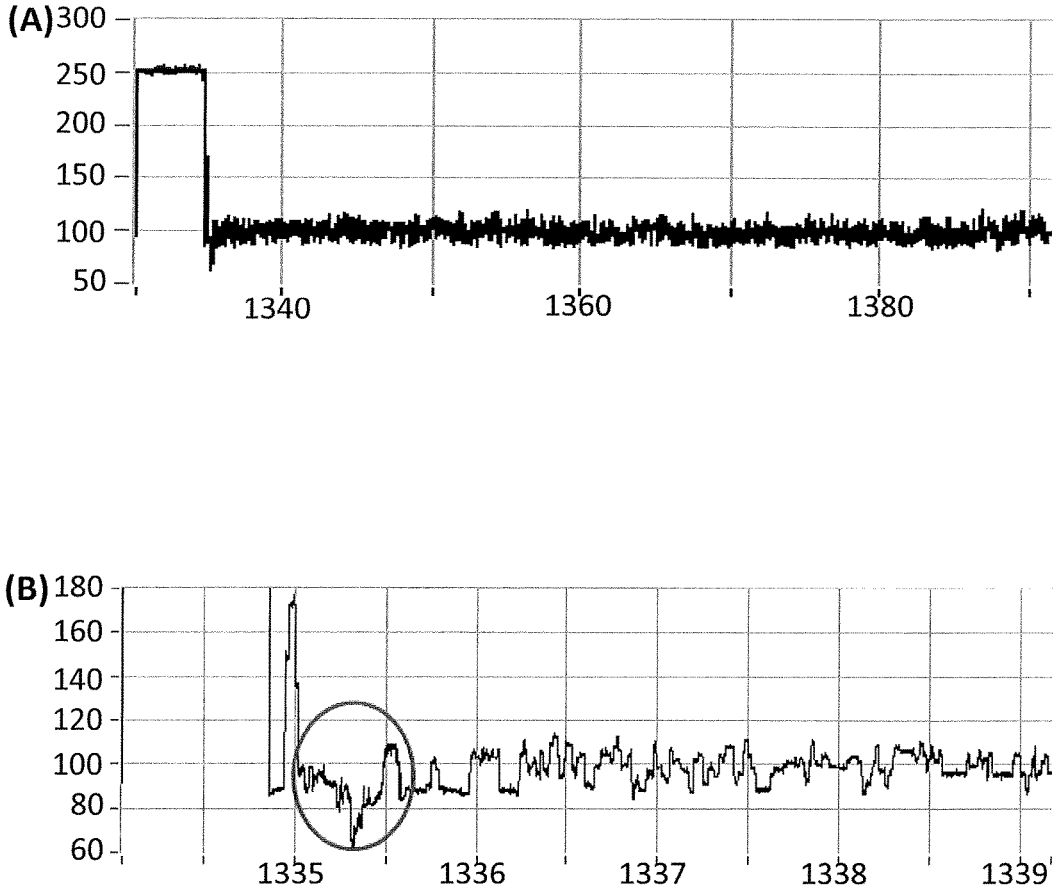

(A)
300
250
200
150
100
50
1340
1360
1380
(B)
180
160
140
120
100
80
60
1335
1336
1337
1338
1339

Figure 14

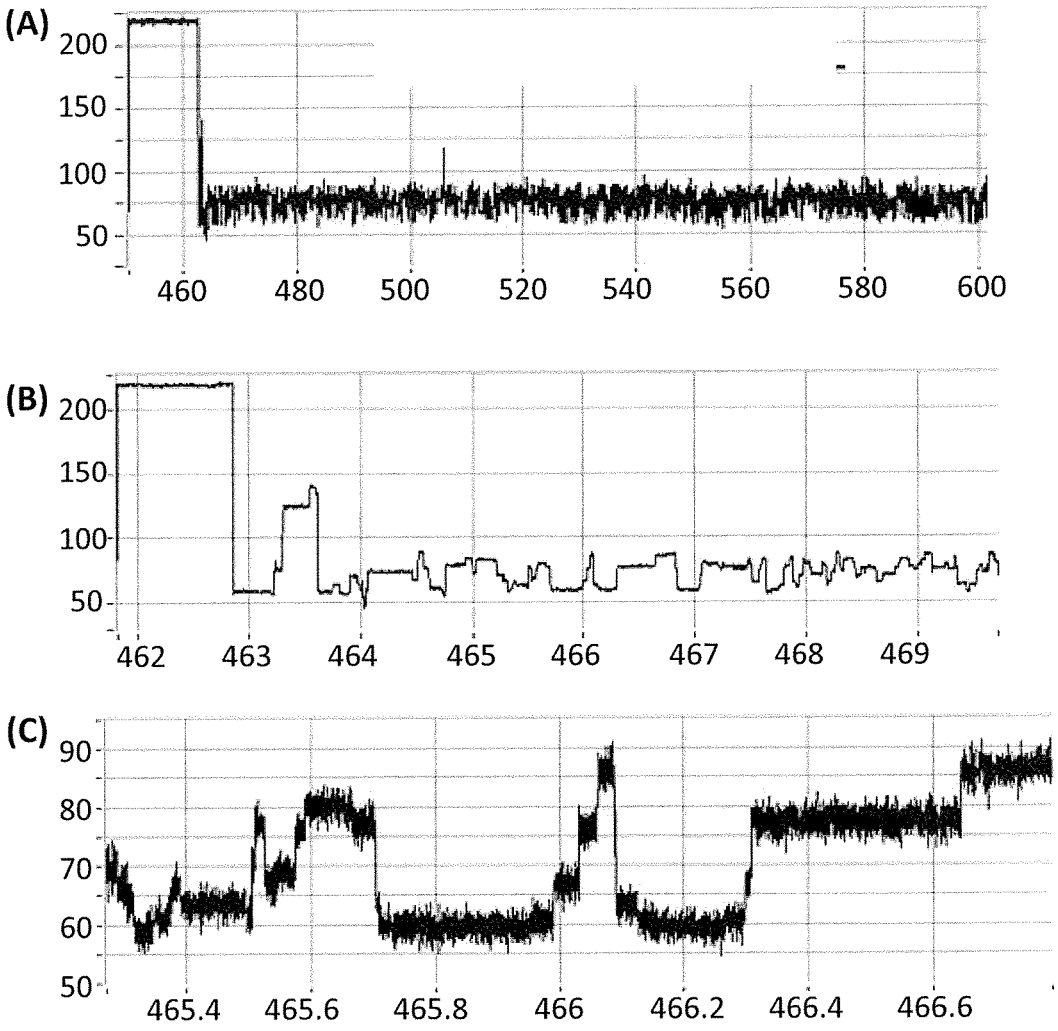

(A)
200
150
100
50
460
480
500
520
540
560
580
600
(B)
200
150
100
50
462
463
464
465
466
467
468
469
(C)
90
80
70
60
50
465.4
465.6
465.8
466
466.2
466.4
466.6

Figure 15

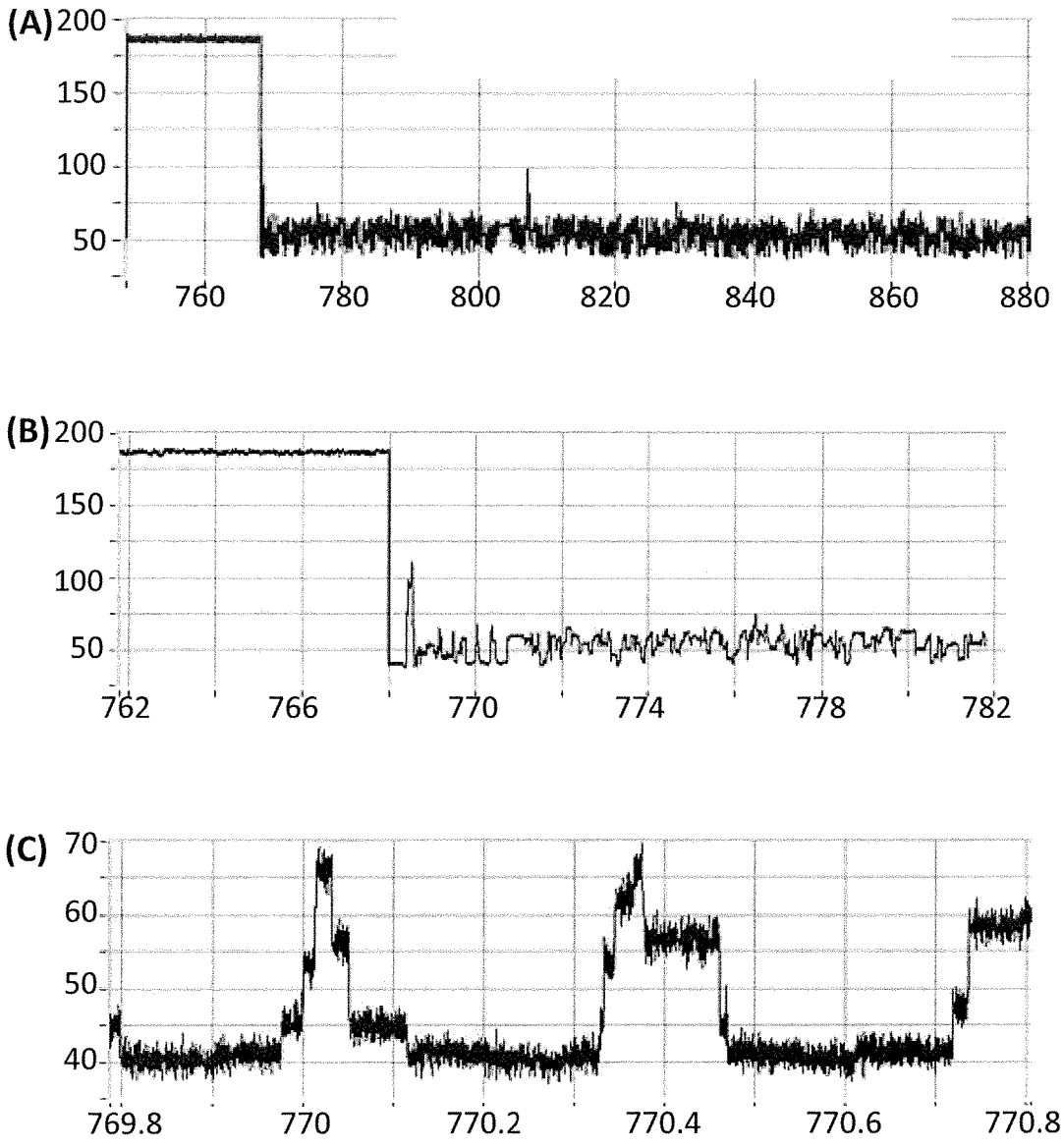

(A)
200
150
100
50
760
780
800
820
840
860
880
(B)
200
150
100
50
762
766
770
774
778
782
(C)
70
60
50
40
769.8
770
770.2
770.4
770.6
770.8

Figure 16

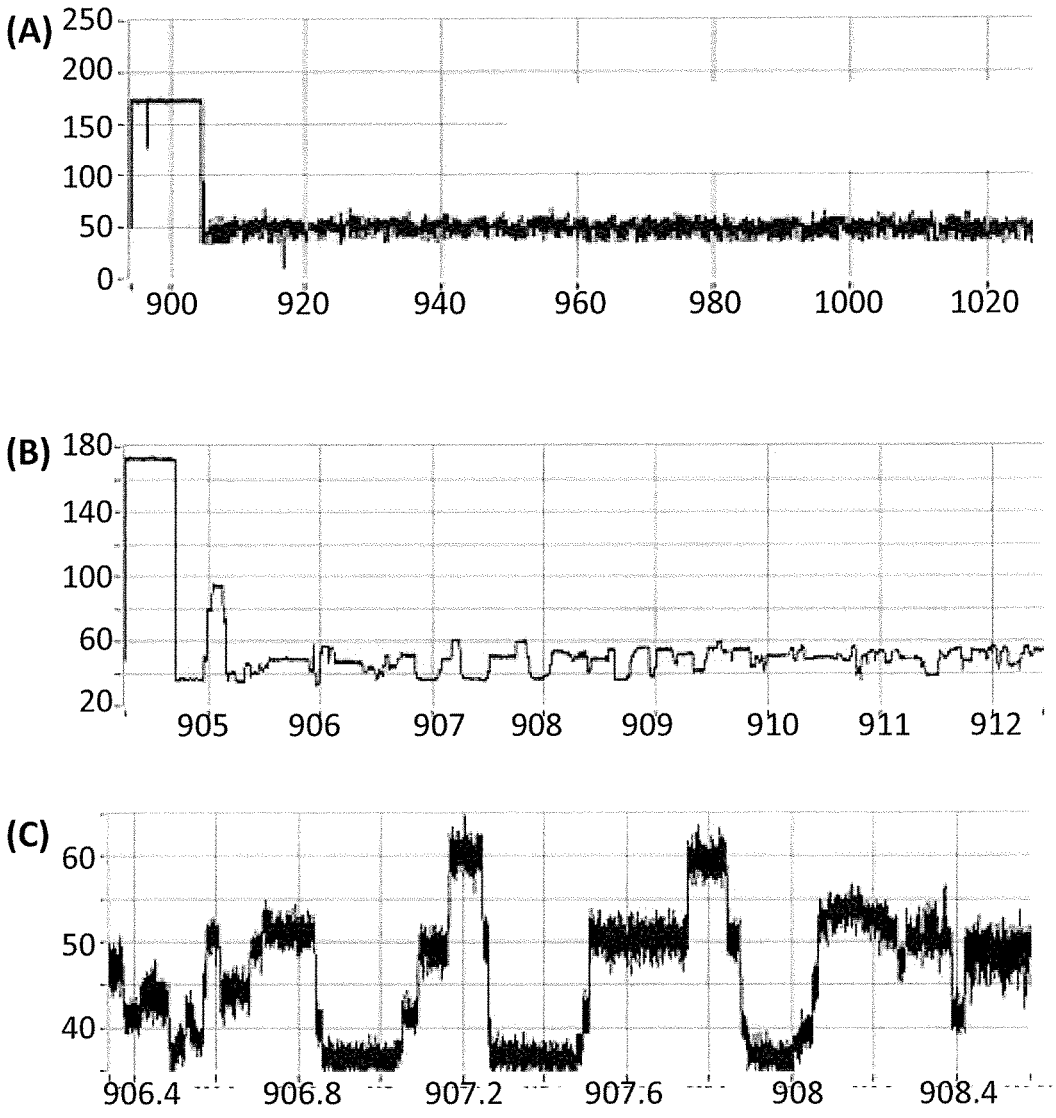

(A)
250
200
150
100
50
0
900
920
940
960
980
1000
1020
(B)
180
140
100
60
20
905
906
907
908
909
910
911
912
(C)
60
50
40
906.4
906.8
907.2
907.6
908
908.4

Figure 17

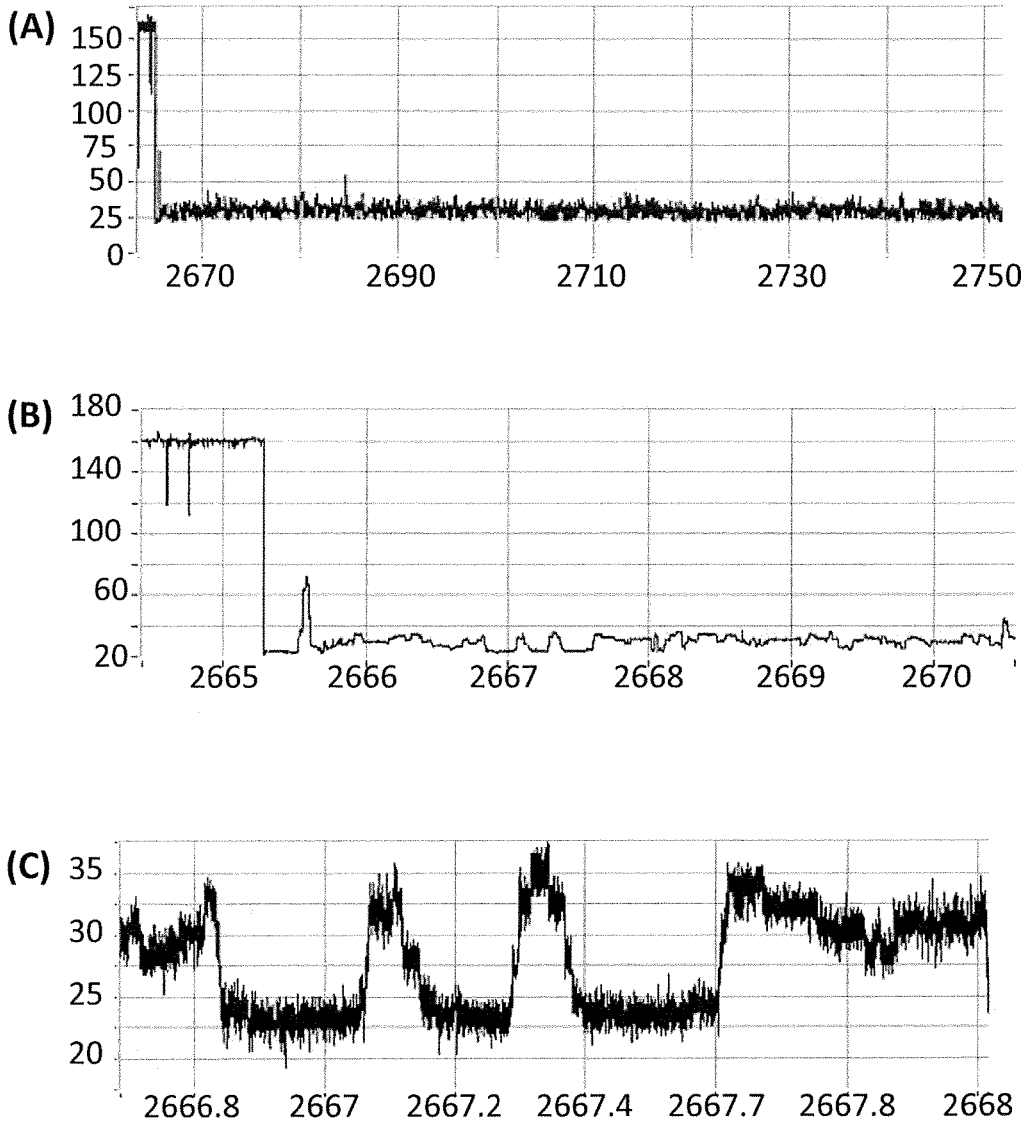

(A)
150
125
100
75
50
25
0
2670
2690
2710
2730
2750
(B)
180
140
100
60
20
2665
2666
2667
2668
2669
2670
(C)
35
30
25
20
2666.8
2667
2667.2
2667.4
2667.7
2667.8
2668

Figure 18

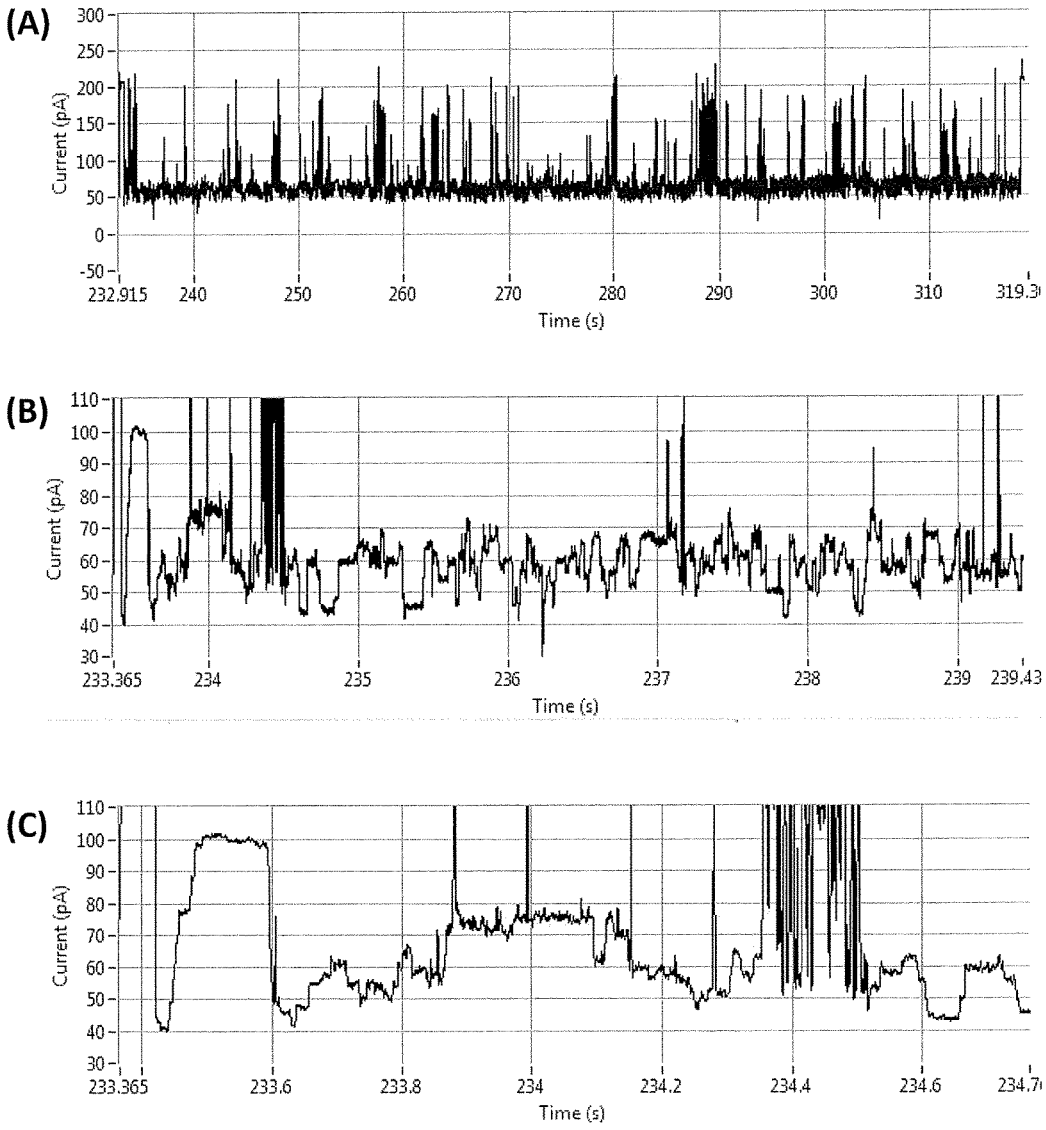

(A)
Current (pA)
300
250
200
150
100
50
0
-50
232.915
240
250
260
270
280
290
300
310
319.3
Time (s)
(B)
Current (pA)
110
100
90
80
70
60
50
40
30
233.365
234
235
236
237
238
239
239.43
Time (s)
(C)
Current (pA)
110
100
90
80
70
60
50
40
30
233.365
233.6
233.8
234
234.2
234.4
234.6
234.7
Time (s)

Figure 19

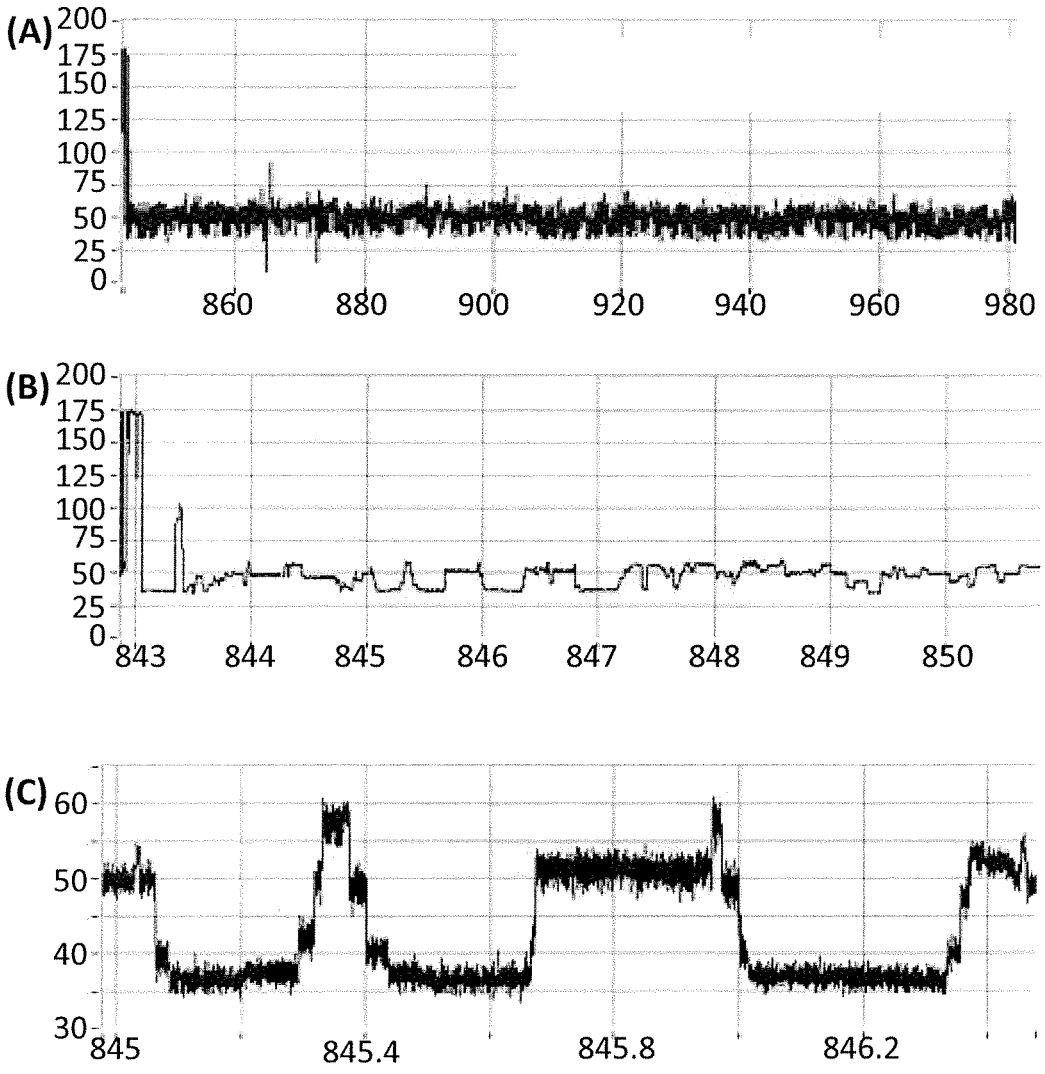
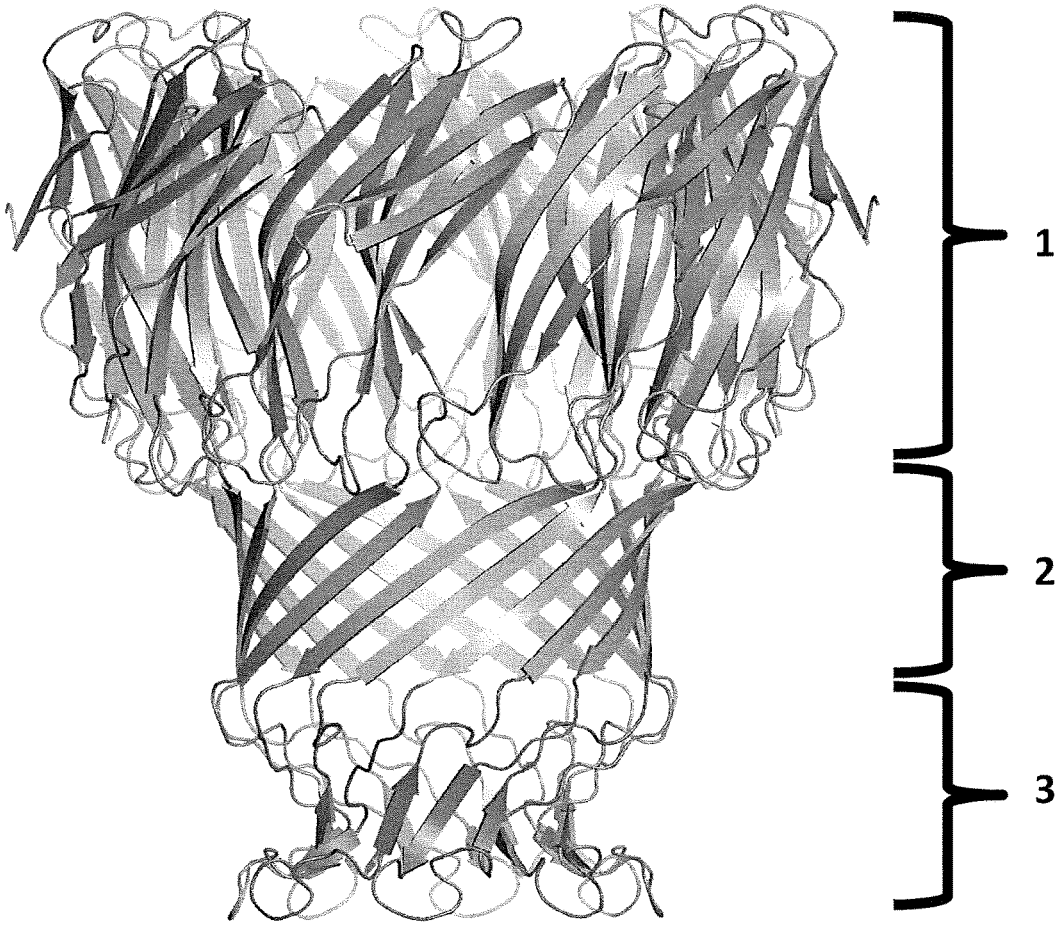

(A)
200
175
150
125
100
75
50
25
0
860
880
900
920
940
960
980
(B)
200
175
150
125
100
75
50
25
0
843
844
845
846
847
848
849
850
(C)
60
50
40
30
845
845.4
845.8
846.2

Figure 20



MUTANT PORES

FIELD OF THE INVENTION

[0001] The invention relates to mutant forms of Msp. The invention also relates to polynucleotide characterisation using Msp.

BACKGROUND OF THE INVENTION

[0002] There is currently a need for rapid and cheap polynucleotide (e.g. DNA or RNA) sequencing and identification technologies across a wide range of applications. Existing technologies are slow and expensive mainly because they rely on amplification techniques to produce large volumes of polynucleotide and require a high quantity of specialist fluorescent chemicals for signal detection.

[0003] Transmembrane pores (nanopores) have great potential as direct, electrical biosensors for polymers and a variety of small molecules. In particular, recent focus has been given to nanopores as a potential DNA sequencing technology.

[0004] When a potential is applied across a nanopore, there is a change in the current flow when an analyte, such as a nucleotide, resides transiently in the barrel for a certain period of time. Nanopore detection of the nucleotide gives a current change of known signature and duration. In the strand sequencing method, a single polynucleotide strand is passed through the pore and the identities of the nucleotides are derived. Strand sequencing can involve the use of a polynucleotide binding protein to control the movement of the polynucleotide through the pore.

[0005] The different forms of Msp are porins from *Mycobacterium smegmatis*. MspA is a 157 kDa octameric porin from *Mycobacterium smegmatis*. Wild-type MspA does not interact with DNA in a manner that allows the DNA to be characterised or sequenced. The structure of MspA and the modifications required for it to interact with and characterise DNA have been well documented (Butler, 2007, Nanopore Analysis of Nucleic Acids, Doctor of Philosophy Dissertation, University of Washington; Gundlach, Proc Natl Acad Sci USA. 2010 Sep. 14; 107(37):16060-5. Epub 2010 Aug. 26; and International Application No. PCT/GB2012/050301 (published as WO/2012/107778). Some key residues have been identified and modified to provide observable interactions between DNA and MspA, which are essential for DNA characterisation or sequencing. In particular, Butler supra teaches that removal of negative charge from all three of positions 90, 91 and 93 is required to provide observable interactions between DNA and MspA.

SUMMARY OF THE INVENTION

[0006] The inventors have surprisingly demonstrated that novel mutants of MspA which retain a negative charge at position 93 may be used to characterise polynucleotides, such as DNA, as long as they have a decreased net negative charge in the inner lining of the cap forming region and/or the barrel forming region. The negative charge at position 93 does not affect the ability of the mutants to capture polynucleotides, interact with polynucleotides and discriminate between the nucleotides in polynucleotides.

[0007] Accordingly, the invention provides a mutant Msp monomer comprising a variant of the sequence shown in SEQ ID NO: 2, wherein the variant:

[0008] (a) does not comprise aspartic acid (D) at position 90;

[0009] (b) does not comprise aspartic acid (D) at position 91;

[0010] (c) comprises aspartic acid (D) or glutamic acid (E) at position 93; and

[0011] (d) comprises one or more modifications which decrease the net negative charge of the inward facing amino acids in the cap forming region and/or the barrel forming region of the monomer.

[0012] The invention also provides:

[0013] A construct comprising two or more covalently attached MspA monomers, wherein at least one of the monomers is a mutant monomer of the invention.

[0014] A polynucleotide which encodes a mutant monomer of the invention or a construct of the invention.

[0015] A homo-oligomeric pore derived from Msp comprising identical mutant monomers of the invention or identical constructs of the invention.

[0016] A hetero-oligomeric pore derived from Msp comprising at least one mutant monomer of the invention or at least one construct of the invention.

[0017] A method of characterising a target polynucleotide, comprising:

[0018] a) contacting the polynucleotide with a pore of the invention such that the polynucleotide moves through the pore; and

[0019] b) taking one or more measurements as the polynucleotide moves with respect to the pore, wherein the measurements are indicative of one or more characteristics of the polynucleotide, and thereby characterising the target polynucleotide.

[0020] A kit for characterising a target polynucleotide comprising (a) a pore of the invention and (b) the components of a membrane.

[0021] An apparatus for characterising target polynucleotides in a sample, comprising (a) a plurality of pores of the invention and (b) a plurality of membranes.

[0022] A method of characterising a target polynucleotide, comprising:

[0023] a) contacting the polynucleotide with a pore of the invention, a polymerase and labelled nucleotides such that phosphate labelled species are sequentially released when nucleotides are added to the first polynucleotide analyte by the polymerase, wherein the phosphate species contain a label specific for each nucleotide; and

[0024] b) detecting the phosphate labelled species using the pore and thereby characterising the polynucleotide.

DESCRIPTION OF THE FIGURES

[0025] FIG. **1** shows DNA construct X which was used in Example 1. Section a of DNA construct X corresponds to SEQ ID NO: 26. Section b corresponds to four iSpC3 spacers. C corresponds to the helicase enzyme T4 Dda—E94C/C109A/C136A/A360C (SEQ ID NO: 24 with mutations E94C/C109A/C136A/A360C) which can bind to the section labelled a. Section d corresponds to SEQ ID NO: 27. Section e corresponds to four 5'-nitroindoles. Section f corresponds to SEQ ID NO: 28. Section g corresponds to SEQ ID NO: 29 which is attached at its 3' end to six iSp18 spacers which are attached at the opposite end to two thymines and 3' cholesterol TEG (labelled h).

[0026] FIG. **2** shows DNA construct Y which was used in Example 1. Section i of DNA construct X corresponds to 25 iSpC3 spacers, which are attached to the 5' end of SEQ ID NO: 3 (labelled j). Section j is the region of construct Y to which the helicase enzyme T4 Dda—E94C/C109A/C136A/A360C can bind (labelled c). Section k corresponds to four iSp18 spacers. Section d corresponds to SEQ ID NO: 27. Section e corresponds to four 5'-nitroindoles. Section 1 corresponds to SEQ ID NO: 4. Section g corresponds to SEQ ID NO: 29 which is attached at its 3' end to six iSp18 spacers which are attached at the opposite end to two thymines and 3' cholesterol TEG (labelled h).

[0027] FIG. **3** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0028] FIG. **4** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for both traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct Y through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/A96D/D118R/Q126R/D134R/E139K)8. Sections B shows a zoomed in region of current trace A.

[0029] FIG. **5** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/N102G/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0030] FIG. **6** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/S103A/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0031] FIG. **7** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/N108S/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0032] FIG. **8** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/N108P/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0033] FIG. **9** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/A96D/N108P/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0034] FIG. **10** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/A96D/N108A/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0035] FIG. **11** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for both traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77 S/L88N/I89F/D90N/D91N/D118R/Q126R/D134R/E139K)8. Sections B shows a zoomed in region of current trace A.

[0036] FIG. **12** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/D90N/D91N/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0037] FIG. **13** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for both traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88K/D90N/D91N/I105E/D118R/Q126R/D134R/E139K)8. Sections B shows a zoomed in region of current trace A.

[0038] FIG. **14** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/D118G/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0039] FIG. **15** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/D118N/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0040] FIG. **16** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/D118R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0041] FIG. **17** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88K/D90N/D91N/N108E/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0042] FIG. **18** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct Y through the MspA nanopore MspA—(G75S/G77S/L88N/

D90N/D91N/T95E/P98K/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0043] FIG. **19** shows example current traces (y-axis label=Current (pA), x-axis label=Time (s) for all three traces) of when a helicase (T4 Dda—E94C/C109A/C136A/A360C) controls the translocation of the DNA construct X through the MspA nanopore MspA—(G75S/G77S/L88N/D90N/D91N/D93N/D118R/Q126R/D134R/E139K)8. Sections B and C show zoomed in regions of current trace A.

[0044] FIG. **20** shows a cartoon representation of the wild-type MspA nanopore. Region 1 corresponds to the cap forming region and includes residues 1-72 and 122-184. Region 2 corresponds to the barrel forming region and includes residues 73-82 and 112-121. Region 3 corresponds to the constriction and loops region and includes residues 83-111.

DESCRIPTION OF THE SEQUENCE LISTING

[0045] SEQ ID NO: 1 shows the codon optimised polynucleotide sequence encoding the wild-type MspA monomer. This mutant lacks the signal sequence.

[0046] SEQ ID NO: 2 shows the amino acid sequence of the mature form of the wild-type MspA monomer. This mutant lacks the signal sequence.

[0047] SEQ ID NOs: 3 and 4 shows polynucleotide sequences used in Example 1.

[0048] SEQ ID NOs: 5 to 7 show the amino acid sequences of MspB, C and D.

[0049] SEQ ID NO: 8 shows the polynucleotide sequence encoding the Phi29 DNA polymerase.

[0050] SEQ ID NO: 9 shows the amino acid sequence of the Phi29 DNA polymerase.

[0051] SEQ ID NO: 10 shows the codon optimised polynucleotide sequence derived from the sbcB gene from *E. coli*. It encodes the exonuclease I enzyme (EcoExo I) from *E. coli*.

[0052] SEQ ID NO: 11 shows the amino acid sequence of exonuclease I enzyme (EcoExo I) from *E. coli*.

[0053] SEQ ID NO: 12 shows the codon optimised polynucleotide sequence derived from the xthA gene from *E. coli*. It encodes the exonuclease III enzyme from *E. coli*.

[0054] SEQ ID NO: 13 shows the amino acid sequence of the exonuclease III enzyme from *E. coli*. This enzyme performs distributive digestion of 5' monophosphate nucleosides from one strand of double stranded DNA (dsDNA) in a 3'-5' direction. Enzyme initiation on a strand requires a 5' overhang of approximately 4 nucleotides.

[0055] SEQ ID NO: 14 shows the codon optimised polynucleotide sequence derived from the recJ gene from *T. thermophilus*. It encodes the RecJ enzyme from *T. thermophilus* (TthRecJ-cd).

[0056] SEQ ID NO: 15 shows the amino acid sequence of the RecJ enzyme from *T thermophilus* (TthRecJ-cd). This enzyme performs processive digestion of 5' monophosphate nucleosides from ssDNA in a 5'-3' direction. Enzyme initiation on a strand requires at least 4 nucleotides.

[0057] SEQ ID NO: 16 shows the codon optimised polynucleotide sequence derived from the bacteriophage lambda exo (redX) gene. It encodes the bacteriophage lambda exonuclease.

[0058] SEQ ID NO: 17 shows the amino acid sequence of the bacteriophage lambda exonuclease. The sequence is one of three identical subunits that assemble into a trimer. The enzyme performs highly processive digestion of nucleotides from one strand of dsDNA, in a 5'-3' direction (http://www.neb.com/nebecomm/products/productM0262.asp). Enzyme initiation on a strand preferentially requires a 5' overhang of approximately 4 nucleotides with a 5' phosphate.

[0059] SEQ ID NO: 18 shows the amino acid sequence of Hel308 Mbu.

[0060] SEQ ID NO: 19 shows the amino acid sequence of Hel308 Csy.

[0061] SEQ ID NO: 20 shows the amino acid sequence of Hel308 Tga.

[0062] SEQ ID NO: 21 shows the amino acid sequence of Hel308 Mhu.

[0063] SEQ ID NO: 22 shows the amino acid sequence of TraI Eco.

[0064] SEQ ID NO: 23 shows the amino acid sequence of XPD Mbu.

[0065] SEQ ID NO: 24 shows the amino acid sequence of Dda 1993.

[0066] SEQ ID NO: 25 shows the amino acid sequence of Trwc Cba.

[0067] SEQ ID NO: 26-29 shows polynucleotide sequences used in Example 1.

DETAILED DESCRIPTION OF THE INVENTION

[0068] It is to be understood that different applications of the disclosed products and methods may be tailored to the specific needs in the art. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments of the invention only, and is not intended to be limiting.

[0069] In addition as used in this specification and the appended claims, the singular forms "a", "an", and "the" include plural referents unless the content clearly dictates otherwise. Thus, for example, reference to "a polynucleotide" includes two or more polynucleotides, reference to "a helicase" includes two or more helicases, reference to "a monomer" refers to two or more monomers, reference to "a pore" includes two or more pores and the like.

[0070] All publications, patents and patent applications cited herein, whether supra or infra, are hereby incorporated by reference in their entirety.

Mutant Msp Monomers

[0071] The present invention provides mutant Msp monomers. The mutant Msp monomers may be used to form the pores of the invention. A mutant Msp monomer is a monomer whose sequence varies from that of a wild-type Msp monomer and which retains the ability to form a pore. Methods for confirming the ability of mutant monomers to form pores are well-known in the art and are discussed in more detail below.

[0072] The mutant monomers have improved polynucleotide reading properties i.e. display improved polynucleotide capture and nucleotide discrimination. In particular, pores constructed from the mutant monomers capture nucleotides and polynucleotides more easily than the wild type. In addition, pores constructed from the mutant monomers display an increased current range, which makes it easier to discriminate between different nucleotides, and a reduced variance of states, which increases the signal-to-noise ratio. In addition, the number of nucleotides contrib-

uting to the current as the polynucleotide moves through pores constructed from the mutants is decreased. This makes it easier to identify a direct relationship between the observed current as the polynucleotide moves through the pore and the polynucleotide sequence.

[0073] A mutant monomer of the invention comprises a variant of the sequence shown in SEQ ID NO: 2. SEQ ID NO: 2 is the wild-type MspA monomer. A variant of SEQ ID NO: 2 is a polypeptide that has an amino acid sequence which varies from that of SEQ ID NO: 2 and which retains its ability to form a pore. The ability of a variant to form a pore can be assayed using any method known in the art. For instance, the variant may be inserted into an amphiphilic layer along with other appropriate subunits and its ability to oligomerise to form a pore may be determined. Methods are known in the art for inserting subunits into membranes, such as amphiphilic layers. For example, subunits may be suspended in a purified form in a solution containing a triblock copolymer membrane such that it diffuses to the membrane and is inserted by binding to the membrane and assembling into a functional state. Alternatively, subunits may be directly inserted into the membrane using the "pick and place" method described in M. A. Holden, H. Bayley. J. Am. Chem. Soc. 2005, 127, 6502-6503 and International Application No. PCT/GB2006/001057 (published as WO 2006/100484).

Positions 90 and 91

[0074] In wild-type MspA, amino acids 90 and 91 are both aspartic acid (D). These amino acids in each monomer form part of an inner constriction of the pore (FIG. **20**). The variant does not comprise aspartic acid (D) at position 90. The variant preferably does not comprise aspartic acid (D) or glutamic acid (E) at position 90. The variant preferably does not have a negatively charged amino acid at position 90.

[0075] The variant does not comprise aspartic acid (D) at position 91. The variant preferably does not comprise aspartic acid (D) or glutamic acid (E) at position 91. The variant preferably does not have a negatively charged amino acid at position 91.

[0076] The variant preferably comprises serine (S), glutamine (Q), leucine (L), methionine (M), isoleucine (I), alanine (A), valine (V), glycine (G), phenylalanine (F), tryptophan (W), tyrosine (Y), histidine (H), threonine (T), arginine (R), lysine (K), asparagine (N) or cysteine (C) at position 90 and/or position 91. Any combinations of these amino acids at positions 90 and 91 are envisaged by the invention. The variant preferably comprises asparagine (N) at position 90 and/or position 91. The variant more preferably comprises asparagine (N) at position 90 and position 91. These amino acids are preferably inserted at position 90 and/or 91 by substitution.

Position 93

[0077] In wild-type MspA, amino acid 93 is aspartic acid (D). This amino acid in each monomer also forms part of an inner constriction of the pore (FIG. **20**).

[0078] The variant comprises aspartic acid (D) or glutamic acid (E) at position 93. The variant therefore has a negative charge at position 93. The glutamic acid (E) is preferably introduced by substitution.

Cap Forming Region

[0079] In wild-type MspA, amino acids 1 to 72 and 122 to 184 form the cap of the pore (FIG. **20**). Of these amino acids, V9, Q12, D13, R14, T15, W40, I49, P53, G54, D56, E57, E59, T61, E63, Y66, Q67, 168, F70, P123, 1125, Q126, E127, V128, A129, T130, F131, S132, V133, D134, S136, G137, E139, V144, H148, T150, V151, T152, F163, R165, 1167, S169, T170 and S173 face inwards into the channel of the pore.

Barrel Forming Region

[0080] In wild-type MspA, amino acids 73 to 82 and 112 to 121 form the barrel of the pore (FIG. **20**). Of these amino acids, S73, G75, G77, N79, S81, G112, S114, S116, D118 and G120 face inwards into the channel of the pore. S73, G75, G77, N79, S81 face inwards in the downwards strand and G112, S114, S116, D118 and G120 face inwards in the upwards strand.

Decreased Net Negative Charge

[0081] The variant comprises one or more modifications which decrease the net negative charge of the inward facing amino acids in the cap forming region and/or the barrel forming region of the monomer. The variant preferably comprises two or more modifications which decrease the net negative charge of the inward facing amino acids in the cap forming region and the barrel forming region of the monomer.

[0082] The variant may comprise any number of modifications, such as 1 or more, 2 or more, 3 or more, 4 or more, 5 or more, 6 or more, 7 or more, 8 or more, 9 or more, 10 or more, 15 or more, 20 or more, 30 or more, or 40 or more modifications.

[0083] The net negative charge may be decreased by any means known in the art. The net negative charge is decreased in a manner that does not interfere with the ability of the mutant monomer to form a pore. This can be measured as discussed above.

[0084] The net negative charge of the inward facing amino acids in the cap forming region and/or the barrel forming region is decreased. This means that the inward facing amino acids in the cap forming region and/or the barrel forming region comprise fewer negatively charged amino acids than in SEQ ID NO: 2 and/or comprises more positively charged amino acids than in SEQ ID NO: 2. The one or more modifications may lead to a net positive charge in the inward facing amino acids in the cap forming region and/or the barrel forming region

[0085] The net charge can be measured using methods known in the art. For instance, the isolectric point may be used to define the net charge of the inward facing amino acids in the cap forming region and/or the barrel forming region.

[0086] The one or more modifications are preferably one or more deletions of negatively charged amino acids. Removal of one or more negatively charged amino acids reduces the net negative charge of the inward facing amino acids in the cap forming region and/or barrel forming region. A negatively charged amino acid is an amino acid with a net negative charge. Negatively charged amino acids include, but are not limited to, aspartic acid (D) and glutamic acid (E). Methods for deleting amino acids from proteins, such as MspA monomers, are well known in the art.

[0087] The one or more modifications are preferably one or more substitutions of negatively charged amino acids with one or more positively charged, uncharged, non-polar and/or aromatic amino acids. A positively charged amino acid is an amino acid with a net positive charge. The positively charged amino acid(s) can be naturally-occurring or non-naturally-occurring. The positively charged amino acid(s) may be synthetic or modified. For instance, modified amino acids with a net positive charge may be specifically designed for use in the invention. A number of different types of modification to amino acids are well known in the art.

[0088] Preferred naturally-occurring positively charged amino acids include, but are not limited to, histidine (H), lysine (K) and arginine (R). Any number and combination of H, K and/or R may be substituted for the inward facing amino acids in the cap forming region and/or barrel forming region.

[0089] The uncharged amino acids, non-polar amino acids and/or aromatic amino acids can be naturally occurring or non-naturally-occurring. They may be synthetic or modified. Uncharged amino acids have no net charge. Suitable uncharged amino acids include, but are not limited to, cysteine (C), serine (S), threonine (T), methionine (M), asparagines (N) and glutamine (Q). Non-polar amino acids have non-polar side chains. Suitable non-polar amino acids include, but are not limited to, glycine (G), alanine (A), proline (P), isoleucine (I), leucine (L) and valine (V). Aromatic amino acids have an aromatic side chain. Suitable aromatic amino acids include, but are not limited to, histidine (H), phenylalanine (F), tryptophan (W) and tyrosine (Y). Any number and combination of these amino acids may be substituted into the inward facing amino acids in the cap forming region and/or the barrel forming region.

[0090] The one or more negatively charged amino acids are preferably substituted with alanine (A), valine (V), asparagine (N) or glycine (G). Preferred substitutions include, but are not limited to, substitution of D with A, substitution of D with V, substitution of D with N and substitution of D with G.

[0091] The one or more modifications are preferably one or more introductions of positively charged amino acids. The introduction of positive charge decreases the net negative charge. The one or more positively charged amino acids may be introduced by addition or substitution. Any amino acid may be substituted with a positively charged amino acid. One or more uncharged amino acids, non-polar amino acids and/or aromatic amino acids may be substituted with one or more positively charged amino acids. Any number of positively charged amino acids may be introduced.

[0092] Wild-type MspA comprises a polar glutamine (Q) at position 126. The one or more modifications preferably reduce the net negative charge at position 126. The one or more modifications preferably increase the net positive charge at positions 126. This can be achieved by replacing the polar amino acid at position 126 or an adjacent or a nearby inward facing amino acid with a positively charged amino acid. The variant preferably comprises a positively charged amino acid at position 126. The variant preferably comprises a positively charged amino acid at one or more of positions 123, 125, 127 and 128. The variant may comprise any number and combination of positively charged amino acids at positions 123, 125, 127 and 128. The positively charged amino acid(s) may be introduced by addition or substitution.

[0093] The one or more modifications are preferably one or more introductions of positively charged amino acids which neutralise one or more negatively charged amino acids. The neutralisation of negative charge decreases the net negative charge. The one or more positively charged amino acids may be introduced by addition or substitution. Any amino acid may be substituted with a positively charged amino acid. One or more uncharged amino acids, non-polar amino acids and/or aromatic amino acids may be substituted with one or more positively charged amino acids. Any number of positively charged amino acids may be introduced. The number is typically the same as the number of negatively charged amino acids in the inward facing amino acids in the cap forming region and/or the barrel forming region.

[0094] The one or more positively charged amino acids may be introduced at any position in the cap forming region and/or the barrel forming region as long as they neutralise the negative charge of the one or more negatively charged amino acids. To effectively neutralise the negative charge, there is typically 5 or fewer amino acids in the variant between each positively charged amino acid that is introduced and the negatively charged amino acid it is neutralising. There is preferably 4 or fewer, 3 or fewer or 2 or fewer amino acids in the variant between each positively charged amino acid that is introduced and the negatively charged amino acid it is neutralising. There is more preferably two amino acids in the variant between each positively charged amino acid that is introduced and the negatively charged amino acid it is neutralising. Each positively charged amino acid is most preferably introduced adjacent in the variant to the negatively charged amino acid it is neutralising.

[0095] The one or more positively charged amino acids may be introduced at any position in the inward facing amino acids in the cap forming region and/or the barrel forming region as long as they neutralise the negative charge of the one or more negatively charged amino acids. To effectively neutralise the negative charge, there is typically 5 or fewer inward facing amino acids between each positively charged amino acid that is introduced and the negatively charged amino acid it is neutralising. There is preferably 4 or fewer, 3 or fewer or 2 or fewer inward facing amino acids between each positively charged amino acid that is introduced and the negatively charged amino acid it is neutralising. There is more preferably one inward facing amino acid between each positively charged amino acid that is introduced and the negatively charged amino acid it is neutralising. Each positively charged amino acid is most preferably introduced at the inward facing position adjacent to the negatively charged amino acid it is neutralising.

[0096] Wild-type MspA comprises aspartic acid (D) at positions 118 and 134 and glutamic acid (E) at position 139. Amino acid 118 in each monomer is present within the barrel of the pore (FIG. **20**). The variant preferably comprises a positively charged amino acid at one or more of positions 114, 116, 120, 123, 70, 73, 75, 77 and 79. Positive charges at one or more of these positions neutralise the negative charge at position 118. Positively charged amino acids may present at any number and combination of positions 114, 116, 120, 123, 70, 73, 75, 77 and 79. The amino acids may be introduced by addition or substitution.

[0097] Amino acids 134 and 139 in each monomer are part of the cap (FIG. **20**). The variant comprises a positively charged amino acid at one or more of positions 129, 132, 136, 137, 59, 61 and 63. Positive charges at one or more of these positions neutralise the negative charge at position 134. Positively charged amino acids may present at any number and combination of positions 129, 132, 136, 137, 59, 61 and 63. The amino acids may be introduced by addition or substitution.

[0098] The variant preferably comprises a positively charged amino acid at one or more of positions 137, 138, 141, 143, 45, 47, 49 and 51. Positive charges at one or more of these positions neutralise the negative charge at position 139. Positively charged amino acids may present at any number and combination of positions 137, 138, 141, 143, 45, 47, 49 and 51. The amino acids may be introduced by addition or substitution.

Positions 118, 126, 134 and 139

[0099] The one or more modifications preferably reduce the net negative charge at one or more of positions 118, 126, 134 and 139. The one or more modifications preferably reduce the net negative charge at 118; 126; 134; 139; 118 and 126; 118 and 134; 118 and 139; 126 and 134; 126 and 139; 134 and 139; 118, 126 and 134; 118, 126 and 139; 118, 134 and 139; 126, 134 and 139; or 118, 126, 134 and 139.

[0100] The variant preferably does not comprise aspartic acid (D) or glutamic acid (E) at one or more of positions 118, 126, 134 and 139. The variant preferably does not comprise aspartic acid (D) or glutamic acid (E) at any of the combination of positions 118, 126, 134 and 139 disclosed above. The variant more preferably comprises arginine (R), glycine (G) or asparagine (N) at one or more of positions 118, 126, 134 and 139, such as any of the combinations of positions 118, 126, 134 and 139 disclosed above. The variant most preferably comprises D118R, Q126R, D134R and E139K.

[0101] Methods for introducing or substituting naturally-occurring amino acids are well known in the art. For instance, methionine (M) may be substituted with arginine (R) by replacing the codon for methionine (ATG) with a codon for arginine (CGT) at the relevant position in a polynucleotide encoding the mutant monomer. The polynucleotide can then be expressed as discussed below.

[0102] Methods for introducing or substituting non-naturally-occurring amino acids are also well known in the art. For instance, non-naturally-occurring amino acids may be introduced by including synthetic aminoacyl-tRNAs in the IVTT system used to express the mutant monomer. Alternatively, they may be introduced by expressing the mutant monomer in *E. coli* that are auxotrophic for specific amino acids in the presence of synthetic (i.e. non-naturally-occurring) analogues of those specific amino acids. They may also be produced by naked ligation if the mutant monomer is produced using partial peptide synthesis.

[0103] The one or more modifications are preferably one or more chemical modifications of one or more negatively charged amino acids which neutralise their negative charge. For instance, the one or more negatively charged amino acids may be reacted with a carbodiimide.

Other Modifications

[0104] The variant preferably comprises one or more of:

[0105] (a) serine (S) at position 75;

[0106] (b) serine (S) at position 77; and

[0107] (c) asparagine (N) or lysine (K) at position 88.

[0108] The variant may comprise any number and combination of (a) to (c), including (a), (b), (c), (a) and (b), (b) and (c), (a) and (c) and (a), (b) and (c). The variant preferably comprises G75S, G77S and L88N.

[0109] The variant most preferably comprises G75S, G77S, L88N, D90N, D91N, D118R, Q126R, D134R and E139K.

[0110] The variant preferably further comprises one or more of:

[0111] (d) phenylalanine (F) at position 89;

[0112] (e) glutamic acid (E) at position 95 and lysine (K) at position 98;

[0113] (f) aspartic acid (D) at position 96;

[0114] (g) glycine (G) at position 102;

[0115] (h) alanine (A) at position 103; and

[0116] (i) alanine (A), serine (S) or proline (P) at position 108.

[0117] The may comprise any number and combination of (d) to (i).

Variants

[0118] In addition to the specific mutations discussed above, the variant may include other mutations. Over the entire length of the amino acid sequence of SEQ ID NO: 2, a variant will preferably be at least 50% homologous to that sequence based on amino acid identity. More preferably, the variant may be at least 55%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90% and more preferably at least 95%, 97% or 99% homologous based on amino acid identity to the amino acid sequence of SEQ ID NO: 2 over the entire sequence. There may be at least 80%, for example at least 85%, 90% or 95%, amino acid identity over a stretch of 100 or more, for example 125, 150, 175 or 200 or more, contiguous amino acids ("hard homology").

[0119] Standard methods in the art may be used to determine homology. For example the UWGCG Package provides the BESTFIT program which can be used to calculate homology, for example used on its default settings (Devereux et at (1984) *Nucleic Acids Research* 12, p387-395). The PILEUP and BLAST algorithms can be used to calculate homology or line up sequences (such as identifying equivalent residues or corresponding sequences (typically on their default settings)), for example as described in Altschul S. F. (1993) J Mol Evol 36:290-300; Altschul, S. F et at (1990) J Mol Biol 215:403-10. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

[0120] SEQ ID NO: 2 is the mature form of the wild-type MspA monomer. The variant may comprise any of the mutations in the MspB, C or D monomers compared with MspA. The mature forms of MspB, C and D are shown in SEQ ID NOs: 5 to 7. In particular, the variant may comprise the following substitution present in MspB: A138P. The variant may comprise one or more of the following substitutions present in MspC: A96G, N102E and A138P. The variant may comprise one or more of the following mutations present in MspD: Deletion of G1, L2V, E5Q, L8V, D13G, W21A, D22E, K47T, I49H, 168V, D91G, A96Q, N102D, S103T, V1041, S136K and G141A. The variant may comprise combinations of one or more of the mutations and substitutions from Msp B, C and D.

[0121] Amino acid substitutions may be made to the amino acid sequence of SEQ ID NO: 2 in addition to those discussed above, for example up to 1, 2, 3, 4, 5, 10, 20 or 30 substitutions. Conservative substitutions replace amino acids with other amino acids of similar chemical structure, similar chemical properties or similar side-chain volume. The amino acids introduced may have similar polarity, hydrophilicity, hydrophobicity, basicity, acidity, neutrality or charge to the amino acids they replace. Alternatively, the conservative substitution may introduce another amino acid that is aromatic or aliphatic in the place of a pre-existing aromatic or aliphatic amino acid. Conservative amino acid changes are well-known in the art and may be selected in accordance with the properties of the 20 main amino acids as defined in Table 1 below. Where amino acids have similar polarity, this can also be determined by reference to the hydropathy scale for amino acid side chains in Table 1.

TABLE 1

Chemical properties of amino acids

Ala	aliphatic, hydrophobic, neutral	Met	hydrophobic, neutral
Cys	polar, hydrophobic, neutral	Asn	polar, hydrophilic, neutral
Asp	polar, hydrophilic, charged (−)	Pro	hydrophobic, neutral
Glu	polar, hydrophilic, charged (−)	Gln	polar, hydrophilic, neutral
Phe	aromatic, hydrophobic, neutral	Arg	polar, hydrophilic, charged (+)
Gly	aliphatic, neutral	Ser	polar, hydrophilic, neutral
His	aromatic, polar, hydrophilic, charged (+)	Thr	polar, hydrophilic, neutral
Ile	aliphatic, hydrophobic, neutral	Val	aliphatic, hydrophobic, neutral
Lys	polar, hydrophilic, charged(+)	Trp	aromatic, hydrophobic, neutral
Leu	aliphatic, hydrophobic, neutral	Tyr	aromatic, polar, hydrophobic

TABLE 2

Hydropathy scale

Side Chain	Hydropathy
Ile	4.5
Val	4.2
Leu	3.8
Phe	2.8
Cys	2.5
Met	1.9
Ala	1.8
Gly	−0.4
Thr	−0.7
Ser	−0.8
Trp	−0.9
Tyr	−1.3
Pro	−1.6
His	−3.2
Glu	−3.5
Gln	−3.5
Asp	−3.5
Asn	−3.5
Lys	−3.9
Arg	−4.5

[0122] One or more amino acid residues of the amino acid sequence of SEQ ID NO: 2 may additionally be deleted from the polypeptides described above. Up to 1, 2, 3, 4, 5, 10, 20 or 30 residues may be deleted, or more.

[0123] Variants may include fragments of SEQ ID NO: 2. Such fragments retain pore forming activity. Fragments may be at least 50, 100, 150 or 200 amino acids in length. Such fragments may be used to produce the pores. A fragment preferably comprises the pore forming domain of SEQ ID NO: 2. Fragments must include one of residues 88, 90, 91, 105, 118 and 134 of SEQ ID NO: 2. Typically, fragments include all of residues 88, 90, 91, 105, 118 and 134 of SEQ ID NO: 2.

[0124] One or more amino acids may be alternatively or additionally added to the polypeptides described above. An extension may be provided at the amino terminal or carboxy terminal of the amino acid sequence of SEQ ID NO: 2 or polypeptide variant or fragment thereof. The extension may be quite short, for example from 1 to 10 amino acids in length. Alternatively, the extension may be longer, for example up to 50 or 100 amino acids. A carrier protein may be fused to an amino acid sequence according to the invention. Other fusion proteins are discussed in more detail below.

[0125] As discussed above, a variant is a polypeptide that has an amino acid sequence which varies from that of SEQ ID NO: 2 and which retains its ability to form a pore. A variant typically contains the regions of SEQ ID NO: 2 that are responsible for pore formation. The pore forming ability of Msp, which contains a β-barrel, is provided by β-sheets in each subunit. A variant of SEQ ID NO: 2 typically comprises the regions in SEQ ID NO: 2 that form β-sheets. One or more modifications can be made to the regions of SEQ ID NO: 2 that form β-sheets as long as the resulting variant retains its ability to form a pore. A variant of SEQ ID NO: 2 preferably includes one or more modifications, such as substitutions, additions or deletions, within its α-helices and/or loop regions.

[0126] The monomers derived from Msp may be modified to assist their identification or purification, for example by the addition of a streptavidin tag or by the addition of a signal sequence to promote their secretion from a cell where the monomer does not naturally contain such a sequence. Other suitable tags are discussed in more detail below. The monomer may be labelled with a revealing label. The revealing label may be any suitable label which allows the monomer to be detected. Suitable labels are described below.

[0127] The monomer derived from Msp may also be produced using D-amino acids. For instance, the monomer derived from Msp may comprise a mixture of L-amino acids and D-amino acids. This is conventional in the art for producing such proteins or peptides.

[0128] The monomer derived from Msp contains one or more specific modifications to facilitate nucleotide discrimination. The monomer derived from Msp may also contain other non-specific modifications as long as they do not interfere with pore formation. A number of non-specific side chain modifications are known in the art and may be made to the side chains of the monomer derived from Msp. Such modifications include, for example, reductive alkylation of amino acids by reaction with an aldehyde followed by reduction with $NaBH_4$, amidination with methylacetimidate or acylation with acetic anhydride.

[0129] The monomer derived from Msp can be produced using standard methods known in the art. The monomer

derived from Msp may be made synthetically or by recombinant means. For example, the monomer may be synthesized by in vitro translation and transcription (IVTT). Suitable methods for producing pores and monomers are discussed in International Application Nos. PCT/GB09/001690 (published as WO 2010/004273), PCT/GB09/001679 (published as WO 2010/004265) or PCT/GB10/000133 (published as WO 2010/086603).

[0130] In some embodiments, the mutant monomer is chemically modified. The mutant monomer can be chemically modified in any way and at any site. The mutant monomer is preferably chemically modified by attachment of a molecule to one or more cysteines (cysteine linkage), attachment of a molecule to one or more lysines, attachment of a molecule to one or more non-natural amino acids, enzyme modification of an epitope or modification of a terminus. Suitable methods for carrying out such modifications are well-known in the art. The mutant monomer may be chemically modified by the attachment of any molecule. For instance, the mutant monomer may be chemically modified by attachment of a dye or a fluorophore.

[0131] In some embodiments, the mutant monomer is chemically modified with a molecular adaptor that facilitates the interaction between a pore comprising the monomer and a target nucleotide or target polynucleotide sequence. The presence of the adaptor improves the host-guest chemistry of the pore and the nucleotide or polynucleotide sequence and thereby improves the sequencing ability of pores formed from the mutant monomer. The principles of host-guest chemistry are well-known in the art. The adaptor has an effect on the physical or chemical properties of the pore that improves its interaction with the nucleotide or polynucleotide sequence. The adaptor may alter the charge of the barrel or channel of the pore or specifically interact with or bind to the nucleotide or polynucleotide sequence thereby facilitating its interaction with the pore.

[0132] The molecular adaptor is preferably a cyclic molecule, a cyclodextrin, a species that is capable of hybridization, a DNA binder or interchelator, a peptide or peptide analogue, a synthetic polymer, an aromatic planar molecule, a small positively-charged molecule or a small molecule capable of hydrogen-bonding.

[0133] The adaptor may be cyclic. A cyclic adaptor preferably has the same symmetry as the pore. The adaptor preferably has eight-fold symmetry since Msp typically has eight subunits around a central axis. This is discussed in more detail below.

[0134] The adaptor typically interacts with the nucleotide or polynucleotide sequence via host-guest chemistry. The adaptor is typically capable of interacting with the nucleotide or polynucleotide sequence. The adaptor comprises one or more chemical groups that are capable of interacting with the nucleotide or polynucleotide sequence. The one or more chemical groups preferably interact with the nucleotide or polynucleotide sequence by non-covalent interactions, such as hydrophobic interactions, hydrogen bonding, Van der Waal's forces, π-cation interactions and/or electrostatic forces. The one or more chemical groups that are capable of interacting with the nucleotide or polynucleotide sequence are preferably positively charged. The one or more chemical groups that are capable of interacting with the nucleotide or polynucleotide sequence more preferably comprise amino groups. The amino groups can be attached to primary, secondary or tertiary carbon atoms. The adaptor even more preferably comprises a ring of amino groups, such as a ring of 6, 7 or 8 amino groups. The adaptor most preferably comprises a ring of eight amino groups. A ring of protonated amino groups may interact with negatively charged phosphate groups in the nucleotide or polynucleotide sequence.

[0135] The correct positioning of the adaptor within the pore can be facilitated by host-guest chemistry between the adaptor and the pore comprising the mutant monomer. The adaptor preferably comprises one or more chemical groups that are capable of interacting with one or more amino acids in the pore. The adaptor more preferably comprises one or more chemical groups that are capable of interacting with one or more amino acids in the pore via non-covalent interactions, such as hydrophobic interactions, hydrogen bonding, Van der Waal's forces, π-cation interactions and/or electrostatic forces. The chemical groups that are capable of interacting with one or more amino acids in the pore are typically hydroxyls or amines. The hydroxyl groups can be attached to primary, secondary or tertiary carbon atoms. The hydroxyl groups may form hydrogen bonds with uncharged amino acids in the pore. Any adaptor that facilitates the interaction between the pore and the nucleotide or polynucleotide sequence can be used.

[0136] Suitable adaptors include, but are not limited to, cyclodextrins, cyclic peptides and cucurbiturils. The adaptor is preferably a cyclodextrin or a derivative thereof. The cyclodextrin or derivative thereof may be any of those disclosed in Eliseev, A. V., and Schneider, H-J. (1994) *J. Am. Chem. Soc.* 116, 6081-6088. The adaptor is more preferably heptakis-6-amino-β-cyclodextrin (am_7-βCD), 6-monodeoxy-6-monoamino-β-cyclodextrin (am_1-βCD) or heptakis-(6-deoxy-6-guanidino)-cyclodextrin (gu_7-βCD). The guanidino group in gu_7-βCD has a much higher pKa than the primary amines in am_7-βCD and so it more positively charged. This gu_7-βCD adaptor may be used to increase the dwell time of the nucleotide in the pore, to increase the accuracy of the residual current measured, as well as to increase the base detection rate at high temperatures or low data acquisition rates.

[0137] If a succinimidyl 3-(2-pyridyldithio)propionate (SPDP) crosslinker is used as discussed in more detail below, the adaptor is preferably heptakis(6-deoxy-6-amino)-6-N-mono(2-pyridyl)dithiopropanoyl-β-cyclodextrin (am_6amPDP_1-βCD).

[0138] More suitable adaptors include γ-cyclodextrins, which comprise 8 sugar units (and therefore have eight-fold symmetry). The γ-cyclodextrin may contain a linker molecule or may be modified to comprise all or more of the modified sugar units used in the β-cyclodextrin examples discussed above.

[0139] The molecular adaptor is preferably covalently attached to the mutant monomer. The adaptor can be covalently attached to the pore using any method known in the art. The adaptor is typically attached via chemical linkage. If the molecular adaptor is attached via cysteine linkage, the one or more cysteines have preferably been introduced to the mutant by substitution. The mutant monomers of the invention can of course comprise a cysteine residue at one or more of positions 88, 90, 91, 103 and 105. The mutant monomer may be chemically modified by attachment of a molecular adaptor to one or more, such as 2, 3, 4 or 5, of these cysteines. Alternatively, the mutant monomer may be chemically modified by attachment of a molecule to one or more

cysteines introduced at other positions. The molecular adaptor is preferably attached to one or more of positions 90, 91 and 103 of SEQ ID NO: 2.

[0140] The reactivity of cysteine residues may be enhanced by modification of the adjacent residues. For instance, the basic groups of flanking arginine, histidine or lysine residues will change the pKa of the cysteines thiol group to that of the more reactive S^- group. The reactivity of cysteine residues may be protected by thiol protective groups such as dTNB. These may be reacted with one or more cysteine residues of the mutant monomer before a linker is attached. The molecule may be attached directly to the mutant monomer. The molecule is preferably attached to the mutant monomer using a linker, such as a chemical crosslinker or a peptide linker.

[0141] Suitable chemical crosslinkers are well-known in the art. Preferred crosslinkers include 2,5-dioxopyrrolidin-1-yl 3-(pyridin-2-yldisulfanyl)propanoate, 2,5-dioxopyrrolidin-1-yl 4-(pyridin-2-yldisulfanyl)butanoate and 2,5-dioxopyrrolidin-1-yl 8-(pyridin-2-yldisulfanyl)octananoate. The most preferred crosslinker is succinimidyl 3-(2-pyridyldithio)propionate (SPDP). Typically, the molecule is covalently attached to the bifunctional crosslinker before the molecule/crosslinker complex is covalently attached to the mutant monomer but it is also possible to covalently attach the bifunctional crosslinker to the monomer before the bifunctional crosslinker/monomer complex is attached to the molecule.

[0142] The linker is preferably resistant to dithiothreitol (DTT). Suitable linkers include, but are not limited to, iodoacetamide-based and Maleimide-based linkers.

[0143] In other embodiment, the monomer may be attached to a polynucleotide binding protein. This forms a modular sequencing system that may be used in the methods of sequencing of the invention. Polynucleotide binding proteins are discussed below.

[0144] The polynucleotide binding protein is preferably covalently attached to the mutant monomer. The protein can be covalently attached to the monomer using any method known in the art. The monomer and protein may be chemically fused or genetically fused. The monomer and protein are genetically fused if the whole construct is expressed from a single polynucleotide sequence. Genetic fusion of a monomer to a polynucleotide binding protein is discussed in International Application No. PCT/GB09/001679 (published as WO 2010/004265).

[0145] If the polynucleotide binding protein is attached via cysteine linkage, the one or more cysteines have preferably been introduced to the mutant by substitution. The mutant monomers of the invention can of course comprise cysteine residues at one or more of positions 10 to 15, 51 to 60, 136 to 139 and 168 to 172. These positions are present in loop regions which have low conservation amongst homologues indicating that mutations or insertions may be tolerated. They are therefore suitable for attaching a polynucleotide binding protein. The reactivity of cysteine residues may be enhanced by modification as described above.

[0146] The polynucleotide binding protein may be attached directly to the mutant monomer or via one or more linkers. The molecule may be attached to the mutant monomer using the hybridization linkers described in International Application No. PCT/GB10/000132 (published as WO 2010/086602). Alternatively, peptide linkers may be used. Peptide linkers are amino acid sequences. The length, flexibility and hydrophilicity of the peptide linker are typically designed such that it does not to disturb the functions of the monomer and molecule. Preferred flexible peptide linkers are stretches of 2 to 20, such as 4, 6, 8, 10 or 16, serine and/or glycine amino acids. More preferred flexible linkers include $(SG)_1$, $(SG)_2$, $(SG)_3$, $(SG)_4$, $(SG)_5$ and $(SG)_8$ wherein S is serine and G is glycine. Preferred rigid linkers are stretches of 2 to 30, such as 4, 6, 8, 16 or 24, proline amino acids. More preferred rigid linkers include $(P)_{12}$ wherein P is proline.

[0147] The mutant monomer may be chemically modified with a molecular adaptor and a polynucleotide binding protein.

[0148] The reactivity of cysteine residues may be enhanced by modification of the adjacent residues. For instance, the basic groups of flanking arginine, histidine or lysine residues will change the pKa of the cysteines thiol group to that of the more reactive S^- group. The reactivity of cysteine residues may be protected by thiol protective groups such as dTNB. These may be reacted with one or more cysteine residues of the monomer before a linker is attached.

[0149] The molecule (with which the monomer is chemically modified) may be attached directly to the monomer or attached via a linker as disclosed in International Application Nos. PCT/GB09/001690 (published as WO 2010/004273), PCT/GB09/001679 (published as WO 2010/004265) or PCT/GB10/000133 (published as WO 2010/086603).

[0150] Any of the proteins described herein, such as the mutant monomers and pores of the invention, may be modified to assist their identification or purification, for example by the addition of histidine residues (a his tag), aspartic acid residues (an asp tag), a streptavidin tag, a flag tag, a SUMO tag, a GST tag or a MBP tag, or by the addition of a signal sequence to promote their secretion from a cell where the polypeptide does not naturally contain such a sequence. An alternative to introducing a genetic tag is to chemically react a tag onto a native or engineered position on the protein. An example of this would be to react a gel-shift reagent to a cysteine engineered on the outside of the protein. This has been demonstrated as a method for separating hemolysin hetero-oligomers (Chem Biol. 1997 July; 4(7):497-505).

[0151] Any of the proteins described herein, such as the mutant monomers and pores of the invention, may be labelled with a revealing label. The revealing label may be any suitable label which allows the protein to be detected. Suitable labels include, but are not limited to, fluorescent molecules, radioisotopes e.g. ^{125}I, ^{35}S, enzymes, antibodies, antigens, polynucleotides and ligands such as biotin.

[0152] Any of the proteins described herein, such as the monomers or pores of the invention, may be made synthetically or by recombinant means. For example, the protein may be synthesized by in vitro translation and transcription (IVTT). The amino acid sequence of the protein may be modified to include non-naturally occurring amino acids or to increase the stability of the protein. When a protein is produced by synthetic means, such amino acids may be introduced during production. The protein may also be altered following either synthetic or recombinant production.

[0153] Proteins may also be produced using D-amino acids. For instance, the protein may comprise a mixture of

L-amino acids and D-amino acids. This is conventional in the art for producing such proteins or peptides.

[0154] The protein may also contain other non-specific modifications as long as they do not interfere with the function of the protein. A number of non-specific side chain modifications are known in the art and may be made to the side chains of the protein(s). Such modifications include, for example, reductive alkylation of amino acids by reaction with an aldehyde followed by reduction with $NaBH_4$, amidination with methylacetimidate or acylation with acetic anhydride.

[0155] Any of the proteins described herein, including the monomers and pores of the invention, can be produced using standard methods known in the art. Polynucleotide sequences encoding a protein may be derived and replicated using standard methods in the art. Polynucleotide sequences encoding a protein may be expressed in a bacterial host cell using standard techniques in the art. The protein may be produced in a cell by in situ expression of the polypeptide from a recombinant expression vector. The expression vector optionally carries an inducible promoter to control the expression of the polypeptide. These methods are described in Sambrook, J. and Russell, D. (2001). Molecular Cloning: A Laboratory Manual, 3rd Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

[0156] Proteins may be produced in large scale following purification by any protein liquid chromatography system from protein producing organisms or after recombinant expression. Typical protein liquid chromatography systems include FPLC, AKTA systems, the Bio-Cad system, the Bio-Rad BioLogic system and the Gilson HPLC system.

Constructs

[0157] The invention also provides a construct comprising two or more covalently attached MspA monomers, wherein at least one of the monomers is a mutant monomer of the invention. The construct of the invention retains its ability to form a pore. This may be determined as discussed above. One or more constructs of the invention may be used to form pores for characterising, such as sequencing, polynucleotides. The construct may comprise at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9 or at least 10 monomers. The construct preferably comprises two monomers. The two or more monomers may be the same or different.

[0158] The construct may comprise one or more monomers which are not mutant monomers of the invention. MspA mutant monomers which are non mutant monomers of the invention comprise comparative variants of SEQ ID NO: 2. At least one monomer in the construct may comprise a comparative variant of the sequence shown in SEQ ID NO: 2 which comprises D90N, D91N, D93N, D118R, D134R and E139K. At least one monomer in the construct may be any of the monomers disclosed in International Application No. PCT/GB2012/050301 (published as WO/2012/107778), including those comprising a comparative variant of the sequence shown in SEQ ID NO: 2 which comprises G75S, G77S, L88N, D90N, D91N, D93N, D118R, Q126R, D134R and E139K. A comparative variant of SEQ ID NO: 2 is at least 50% homologous to SEQ ID NO: 2 over its entire sequence based on amino acid identity. More preferably, the comparative variant may be at least 55%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90% and more preferably at least 95%, 97% or 99% homologous based on amino acid identity to the amino acid sequence of SEQ ID NO: 2 over the entire sequence.

[0159] At least one monomer in the construct is a mutant monomer of the invention. All of the monomers in the construct may be a mutant monomer of the invention. The mutant monomers may be the same or different. In a more preferred embodiment, the construct comprises two monomers of the invention.

[0160] The monomers in the construct are preferably genetically fused. Monomers are genetically fused if the whole construct is expressed from a single polynucleotide sequence. The coding sequences of the monomers may be combined in any way to form a single polynucleotide sequence encoding the construct.

[0161] The monomers may be genetically fused in any configuration. The monomers may be fused via their terminal amino acids. For instance, the amino terminus of the one monomer may be fused to the carboxy terminus of another monomer. The second and subsequent monomers in the construct (in the amino to carboxy direction) may comprise a methionine at their amino terminal ends (each of which is fused to the carboxy terminus of the previous monomer). For instance, if M is a monomer (without an amino terminal methionine) and mM is a monomer with an amino terminal methionine, the construct may comprise the sequence M-mM, M-mM-mM or M-mM-mM-mM. The presences of these methionines typically results from the expression of the start codons (i.e. ATGs) at the 5' end of the polynucleotides encoding the second or subsequent monomers within the polynucleotide encoding entire construct. The first monomer in the construct (in the amino to carboxy direction) may also comprise a methionine (e.g. mM-mM, mM-mM-mM or mM-mM-mM-mM).

[0162] The two or more monomers may be genetically fused directly together. The monomers are preferably genetically fused using a linker. The linker may be designed to constrain the mobility of the monomers. Preferred linkers are amino acid sequences (i.e. peptide linkers). Any of the peptide linkers discussed above may be used.

[0163] In another preferred embodiment, the monomers are chemically fused. Two monomers are chemically fused if the two parts are chemically attached, for instance via a chemical crosslinker. Any of the chemical crosslinkers discussed above may be used. The linker may be attached to one or more cysteine residues introduced into a mutant monomer of the invention. Alternatively, the linker may be attached to a terminus of one of the monomers in the construct.

[0164] If a construct contains different monomers, crosslinkage of monomers to themselves may be prevented by keeping the concentration of linker in a vast excess of the monomers. Alternatively, a "lock and key" arrangement may be used in which two linkers are used. Only one end of each linker may react together to form a longer linker and the other ends of the linker each react with a different monomers. Such linkers are described in International Application No. PCT/GB10/000132 (published as WO 2010/086602).

Polynucleotides

[0165] The present invention also provides polynucleotide sequences which encode a mutant monomer of the invention. The mutant monomer may be any of those discussed above. The polynucleotide sequence preferably comprises a sequence at least 50%, 60%, 70%, 80%, 90% or 95%

homologous based on nucleotide identity to the sequence of SEQ ID NO: 1 over the entire sequence. There may be at least 80%, for example at least 85%, 90% or 95% nucleotide identity over a stretch of 300 or more, for example 375, 450, 525 or 600 or more, contiguous nucleotides ("hard homology"). Homology may be calculated as described above. The polynucleotide sequence may comprise a sequence that differs from SEQ ID NO: 1 on the basis of the degeneracy of the genetic code.

[0166] The present invention also provides polynucleotide sequences which encode any of the genetically fused constructs of the invention. The polynucleotide preferably comprises two or more variants of the sequence shown in SEQ ID NO: 1. The polynucleotide sequence preferably comprises two or more sequences having at least 50%, 60%, 70%, 80%, 90% or 95% homology to SEQ ID NO: 1 based on nucleotide identity over the entire sequence. There may be at least 80%, for example at least 85%, 90% or 95% nucleotide identity over a stretch of 600 or more, for example 750, 900, 1050 or 1200 or more, contiguous nucleotides ("hard homology"). Homology may be calculated as described above.

[0167] Polynucleotide sequences may be derived and replicated using standard methods in the art. Chromosomal DNA encoding wild-type Msp may be extracted from a pore producing organism, such as *Mycobacterium smegmatis*. The gene encoding the pore subunit may be amplified using PCR involving specific primers. The amplified sequence may then undergo site-directed mutagenesis. Suitable methods of site-directed mutagenesis are known in the art and include, for example, combine chain reaction. Polynucleotides encoding a construct of the invention can be made using well-known techniques, such as those described in Sambrook, J. and Russell, D. (2001). Molecular Cloning: A Laboratory Manual, 3rd Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

[0168] The resulting polynucleotide sequence may then be incorporated into a recombinant replicable vector such as a cloning vector. The vector may be used to replicate the polynucleotide in a compatible host cell. Thus polynucleotide sequences may be made by introducing a polynucleotide into a replicable vector, introducing the vector into a compatible host cell, and growing the host cell under conditions which bring about replication of the vector. The vector may be recovered from the host cell. Suitable host cells for cloning of polynucleotides are known in the art and described in more detail below.

[0169] The polynucleotide sequence may be cloned into suitable expression vector. In an expression vector, the polynucleotide sequence is typically operably linked to a control sequence which is capable of providing for the expression of the coding sequence by the host cell. Such expression vectors can be used to express a pore subunit.

[0170] The term "operably linked" refers to a juxtaposition wherein the components described are in a relationship permitting them to function in their intended manner. A control sequence "operably linked" to a coding sequence is ligated in such a way that expression of the coding sequence is achieved under conditions compatible with the control sequences. Multiple copies of the same or different polynucleotide sequences may be introduced into the vector.

[0171] The expression vector may then be introduced into a suitable host cell. Thus, a mutant monomer or construct of the invention can be produced by inserting a polynucleotide sequence into an expression vector, introducing the vector into a compatible bacterial host cell, and growing the host cell under conditions which bring about expression of the polynucleotide sequence. The recombinantly-expressed monomer or construct may self-assemble into a pore in the host cell membrane. Alternatively, the recombinant pore produced in this manner may be removed from the host cell and inserted into another membrane. When producing pores comprising at least two different monomers or constructs, the different monomers or constructs may be expressed separately in different host cells as described above, removed from the host cells and assembled into a pore in a separate membrane, such as a rabbit cell membrane.

[0172] The vectors may be for example, plasmid, virus or phage vectors provided with an origin of replication, optionally a promoter for the expression of the said polynucleotide sequence and optionally a regulator of the promoter. The vectors may contain one or more selectable marker genes, for example a tetracycline resistance gene. Promoters and other expression regulation signals may be selected to be compatible with the host cell for which the expression vector is designed. A T7, trc, lac, ara or λ_L promoter is typically used.

[0173] The host cell typically expresses the monomer or construct at a high level. Host cells transformed with a polynucleotide sequence will be chosen to be compatible with the expression vector used to transform the cell. The host cell is typically bacterial and preferably *Escherichia coli*. Any cell with a λ DE3 lysogen, for example C41 (DE3), BL21 (DE3), JM109 (DE3), B834 (DE3), TUNER, Origami and Origami B, can express a vector comprising the T7 promoter. In addition to the conditions listed above any of the methods cited in Proc Natl Acad Sci USA. 2008 Dec. 30; 105(52):20647-52 may be used to express the Msp proteins.

[0174] The invention also comprises a method of producing a mutant monomer of the invention or a construct of the invention. The method comprises expressing a polynucleotide of the invention in a suitable host cell. The polynucleotide is preferably part of a vector and is preferably operably linked to a promoter.

Pores

[0175] The invention also provides various pores. The pores of the invention are ideal for characterising, such as sequencing, polynucleotide sequences because they can discriminate between different nucleotides with a high degree of sensitivity. The pores can surprisingly distinguish between the four nucleotides in DNA and RNA. The pores of the invention can even distinguish between methylated and unmethylated nucleotides. The base resolution of pores of the invention is surprisingly high. The pores show almost complete separation of all four DNA nucleotides. The pores further discriminate between deoxycytidine monophosphate (dCMP) and methyl-dCMP based on the dwell time in the pore and the current flowing through the pore.

[0176] The pores of the invention can also discriminate between different nucleotides under a range of conditions. In particular, the pores will discriminate between nucleotides under conditions that are favourable to the characterising, such as sequencing, of nucleic acids. The extent to which the pores of the invention can discriminate between different nucleotides can be controlled by altering the applied potential, the salt concentration, the buffer, the temperature and the presence of additives, such as urea, betaine and DTT.

This allows the function of the pores to be fine-tuned, particularly when sequencing. This is discussed in more detail below. The pores of the invention may also be used to identify polynucleotide polymers from the interaction with one or more monomers rather than on a nucleotide by nucleotide basis.

[0177] A pore of the invention may be isolated, substantially isolated, purified or substantially purified. A pore of the invention is isolated or purified if it is completely free of any other components, such as lipids or other pores. A pore is substantially isolated if it is mixed with carriers or diluents which will not interfere with its intended use. For instance, a pore is substantially isolated or substantially purified if it is present in a form that comprises less than 10%, less than 5%, less than 2% or less than 1% of other components, such as triblock copolymers, lipids or other pores. Alternatively, a pore of the invention may be present in a membrane. Suitable membranes are discussed below.

[0178] A pore of the invention may be present as an individual or single pore. Alternatively, a pore of the invention may be present in a homologous or heterologous population of two or more pores.

Homo-Oligomeric Pores

[0179] The invention also provides a homo-oligomeric pore derived from Msp comprising identical mutant monomers of the invention. The homo-oligomeric pore may comprise any of the mutants of the invention. The homo-oligomeric pore of the invention is ideal for characterising, such as sequencing, polynucleotides. The homo-oligomeric pore of the invention may have any of the advantages discussed above.

[0180] The homo-oligomeric pore may contain any number of mutant monomers. The pore typically comprises at least 7, at least 8, at least 9 or at least 10 identical mutant monomers, such as 7, 8, 9 or 10 mutant monomers. The pore preferably comprises eight identical mutant monomers. One or more, such as 2, 3, 4, 5, 6, 7, 8, 9 or 10, of the mutant monomers is preferably chemically modified as discussed above.

[0181] Methods for making pores are discussed in more detail below.

Hetero-Oligomeric Pores

[0182] The invention also provides a hetero-oligomeric pore derived from Msp comprising at least one mutant monomer of the invention. The hetero-oligomeric pore of the invention is ideal for characterising, such as sequencing, polynucleotides. Hetero-oligomeric pores can be made using methods known in the art (e.g. Protein Sci. 2002 July; 11(7):1813-24).

[0183] The hetero-oligomeric pore contains sufficient monomers to form the pore. The monomers may be of any type. The pore typically comprises at least 7, at least 8, at least 9 or at least 10 monomers, such as 7, 8, 9 or 10 monomers. The pore preferably comprises eight monomers.

[0184] In a preferred embodiment, all of the monomers (such as 10, 9, 8 or 7 of the monomers) are mutant monomers of the invention and at least one of them differs from the others. In a more preferred embodiment, the pore comprises eight mutant monomers of the invention and at least one of them differs from the others. They may all differ from one another.

[0185] In another preferred embodiment, at least one of the mutant monomers is not a mutant monomer of the invention. In this embodiment, the remaining monomers are preferably mutant monomers of the invention. Hence, the pore may comprise 9, 8, 7, 6, 5, 4, 3, 2 or 1 mutant monomers of the invention. Any number of the monomers in the pore may not be a mutant monomer of the invention. The pore preferably comprises seven mutant monomers of the invention and a monomer which is not a monomer of the invention, such as a monomer comprising a comparative variant as discussed above. The mutant monomers of the invention may be the same or different.

[0186] In all the embodiments discussed above, one or more, such as 2, 3, 4, 5, 6, 7, 8, 9 or 10, of the mutant monomers is preferably chemically modified as discussed above.

[0187] Methods for making pores are discussed in more detail below.

Construct-Containing Pores

[0188] The invention also provides a pore comprising at least one construct of the invention. A construct of the invention comprises two or more covalently attached monomers derived from Msp wherein at least one of the monomers is a mutant monomer of the invention. In other words, a construct must contain more than one monomer. The pore contains sufficient constructs and, if necessary, monomers to form the pore. For instance, an octameric pore may comprise (a) four constructs each comprising two constructs, (b) two constructs each comprising four monomers or (b) one construct comprising two monomers and six monomers that do not form part of a construct. Other combinations of constructs and monomers can be envisaged by the skilled person.

[0189] At least two of the monomers in the pore are in the form of a construct of the invention. The construct, and hence the pore, comprises at least one mutant monomer of the invention. The pore typically comprises at least 7, at least 8, at least 9 or at least 10 monomers, such as 7, 8, 9 or 10 monomers, in total (at least two of which must be in a construct). The pore preferably comprises eight monomers (at least two of which must be in a construct).

[0190] The construct containing pore may be a homo-oligomer (i.e. include identical constructs) or be a hetero-oligomer (i.e. where at least one construct differs from the others).

[0191] A pore typically contains (a) one construct comprising two monomers and (b) 5, 6, 7 or 8 monomers. The construct may be any of those discussed above. The monomers may be any of those discussed above, including mutant monomers of the invention and mutant monomers comprising a comparative variant of SEQ ID NO: 2 as discussed above.

[0192] Another typical pore comprises more than one construct of the invention, such as two, three or four constructs of the invention. If necessary, such pores further comprise sufficient additional monomers or constructs to form the pore. The additional monomer(s) may be any of those discussed above, including mutant monomers of the invention and mutant monomers comprising a comparative variant of SEQ ID NO: 2 as discussed above. The additional construct(s) may be any of those discussed above or may be a construct comprising two or more covalently attached

MspA monomers each comprising a comparative variant of SEQ ID NO: 2 as discussed above.

[0193] A further pore of the invention comprises only constructs comprising 2 monomers, for example a pore may comprise 4, 5, 6, 7 or 8 constructs comprising 2 monomers. At least one construct is a construct of the invention, i.e. at least one monomer in the at least one construct, and preferably each monomer in the at least one construct, is a mutant monomer of the invention. All of the constructs comprising 2 monomers may be constructs of the invention.

[0194] A specific pore according to the invention comprises four constructs of the invention each comprising two monomers, wherein at least one monomer in each construct, and preferably each monomer in each construct, is a mutant monomer of the invention. The constructs may oligomerise into a pore with a structure such that only one monomer of each construct contributes to the channel of the pore. Typically the other monomers of the construct will be on the outside of the channel of the pore. For example, pores of the invention may comprise 5, 6, 7 or 8 constructs comprising 2 monomers where the channel comprises 8 monomers.

[0195] Mutations can be introduced into the construct as described above. The mutations may be alternating, i.e. the mutations are different for each monomer within a two monomer construct and the constructs are assembled as a homo-oligomer resulting in alternating modifications. In other words, monomers comprising MutA and MutB are fused and assembled to form an A-B:A-B:A-B:A-B pore. Alternatively, the mutations may be neighbouring, i.e. identical mutations are introduced into two monomers in a construct and this is then oligomerised with different mutant monomers or constructs. In other words, monomers comprising MutA are fused follow by oligomerisation with MutB-containing monomers to form A-A:B:B:B:B:B:B.

[0196] One or more of the monomers of the invention in a construct-containing pore may be chemically-modified as discussed above.

Polynucleotide Characterisation

[0197] The invention provides a method of characterising a target polynucleotide. The method involves measuring one or more characteristics of the target polynucleotide. The target polynucleotide may also be called the template polynucleotide or the polynucleotide of interest.

Polynucleotide

[0198] A polynucleotide, such as a nucleic acid, is a macromolecule comprising two or more nucleotides. The polynucleotide or nucleic acid may comprise any combination of any nucleotides. The nucleotides can be naturally occurring or artificial. One or more nucleotides in the polynucleotide can be oxidized or methylated. One or more nucleotides in the polynucleotide may be damaged. For instance, the polynucleotide may comprise a pyrimidine dimer. Such dimers are typically associated with damage by ultraviolet light and are the primary cause of skin melanomas. One or more nucleotides in the polynucleotide may be modified, for instance with a label or a tag. Suitable labels are described below. The polynucleotide may comprise one or more spacers.

[0199] A nucleotide typically contains a nucleobase, a sugar and at least one phosphate group. The nucleobase and sugar form a nucleoside.

[0200] The nucleobase is typically heterocyclic. Nucleobases include, but are not limited to, purines and pyrimidines and more specifically adenine (A), guanine (G), thymine (T), uracil (U) and cytosine (C).

[0201] The sugar is typically a pentose sugar. Nucleotide sugars include, but are not limited to, ribose and deoxyribose. The sugar is preferably a deoxyribose.

[0202] The polynucleotide preferably comprises the following nucleosides: deoxyadenosine (dA), deoxyuridine (dU) and/or thymidine (dT), deoxyguanosine (dG) and deoxycytidine (dC).

[0203] The nucleotide is typically a ribonucleotide or deoxyribonucleotide. The nucleotide typically contains a monophosphate, diphosphate or triphosphate. The nucleotide may comprise more than three phosphates, such as 4 or 5 phosphates. Phosphates may be attached on the 5' or 3' side of a nucleotide. Nucleotides include, but are not limited to, adenosine monophosphate (AMP), guanosine monophosphate (GMP), thymidine monophosphate (TMP), uridine monophosphate (UMP), 5-methylcytidine monophosphate, 5-hydroxymethylcytidine monophosphate, cytidine monophosphate (CMP), cyclic adenosine monophosphate (cAMP), cyclic guanosine monophosphate (cGMP), deoxyadenosine monophosphate (dAMP), deoxyguanosine monophosphate (dGMP), deoxythymidine monophosphate (dTMP), deoxyuridine monophosphate (dUMP), deoxycytidine monophosphate (dCMP) and deoxymethylcytidine monophosphate. The nucleotides are preferably selected from AMP, TMP, GMP, CMP, UMP, dAMP, dTMP, dGMP, dCMP and dUMP.

[0204] A nucleotide may be abasic (i.e. lack a nucleobase). A nucleotide may also lack a nucleobase and a sugar (i.e. is a C3 spacer).

[0205] The nucleotides in the polynucleotide may be attached to each other in any manner. The nucleotides are typically attached by their sugar and phosphate groups as in nucleic acids. The nucleotides may be connected via their nucleobases as in pyrimidine dimers.

[0206] The polynucleotide may be single stranded or double stranded. At least a portion of the polynucleotide is preferably double stranded.

[0207] The polynucleotide can be a nucleic acid, such as deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). The polynucleotide can comprise one strand of RNA hybridised to one strand of DNA. The polynucleotide may be any synthetic nucleic acid known in the art, such as peptide nucleic acid (PNA), glycerol nucleic acid (GNA), threose nucleic acid (TNA), locked nucleic acid (LNA) or other synthetic polymers with nucleotide side chains. The PNA backbone is composed of repeating N-(2-aminoethyl)-glycine units linked by peptide bonds. The GNA backbone is composed of repeating glycol units linked by phosphodiester bonds. The TNA backbone is composed of repeating threose sugars linked together by phosphodiester bonds. LNA is formed from ribonucleotides as discussed above having an extra bridge connecting the 2' oxygen and 4' carbon in the ribose moiety.

[0208] The polynucleotide is most preferably ribonucleic nucleic acid (RNA) or deoxyribonucleic acid (DNA).

[0209] The polynucleotide can be any length. For example, the polynucleotide can be at least 10, at least 50, at least 100, at least 150, at least 200, at least 250, at least 300, at least 400 or at least 500 nucleotides or nucleotide pairs in length. The polynucleotide can be 1000 or more

nucleotides or nucleotide pairs, 5000 or more nucleotides or nucleotide pairs in length or 100000 or more nucleotides or nucleotide pairs in length.

[0210] Any number of polynucleotides can be investigated. For instance, the method of the invention may concern characterising 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 50, 100 or more polynucleotides. If two or more polynucleotides are characterized, they may be different polynucleotides or two instances of the same polynucleotide.

[0211] The polynucleotide can be naturally occurring or artificial. For instance, the method may be used to verify the sequence of a manufactured oligonucleotide. The method is typically carried out in vitro.

Sample

[0212] Each analyte is typically present in any suitable sample. The invention is typically carried out on two or more samples that are known to contain or suspected to contain the analytes. Alternatively, the invention may be carried out on two or more samples to confirm the identity of two or more analytes whose presence in the samples is known or expected.

[0213] The first sample and/or second sample may be a biological sample. The invention may be carried out in vitro using at least one sample obtained from or extracted from any organism or microorganism. The first sample and/or second sample may be a non-biological sample. The non-biological sample is preferably a fluid sample. Examples of non-biological samples include surgical fluids, water such as drinking water, sea water or river water, and reagents for laboratory tests.

[0214] The first sample and/or second sample is typically processed prior to being used in the invention, for example by centrifugation or by passage through a membrane that filters out unwanted molecules or cells, such as red blood cells. The first sample and/or second sample may be measured immediately upon being taken. The first sample and/or second sample may also be typically stored prior to assay, preferably below −70° C.

Characterisation

[0215] The method may involve measuring two, three, four or five or more characteristics of the polynucleotide. The one or more characteristics are preferably selected from (i) the length of the polynucleotide, (ii) the identity of the polynucleotide, (iii) the sequence of the polynucleotide, (iv) the secondary structure of the polynucleotide and (v) whether or not the polynucleotide is modified. Any combination of (i) to (v) may be measured in accordance with the invention, such as {i}, {ii}, {iii}, {iv}, {v}, {i,ii}, {i,iii}, {i,iv}, {i,v}, {ii,iii}, {ii,iv}, {ii,v}, {iii,iv}, {iii,v}, {iv,v}, {i,ii,iv}, {i,ii,v}, {i,iii,iv}, {i,iii,v}, {i,iv,v}, {ii,iii,iv}, {ii, iii,v}, {ii,iv,v}, {iii,iv,v}, {i,ii,iii,iv}, {i,ii,iii,v}, {i,ii,iv,v}, {i,iii,iv,v}, {ii,iii,iv,v} or {i,ii,iii,iv,v}. Different combinations of (i) to (v) may be measured for the first polynucleotide compared with the second polynucleotide, including any of those combinations listed above.

[0216] For (i), the length of the polynucleotide may be measured for example by determining the number of interactions between the polynucleotide and the pore or the duration of interaction between the polynucleotide and the pore.

[0217] For (ii), the identity of the polynucleotide may be measured in a number of ways. The identity of the polynucleotide may be measured in conjunction with measurement of the sequence of the polynucleotide or without measurement of the sequence of the polynucleotide. The former is straightforward; the polynucleotide is sequenced and thereby identified. The latter may be done in several ways. For instance, the presence of a particular motif in the polynucleotide may be measured (without measuring the remaining sequence of the polynucleotide). Alternatively, the measurement of a particular electrical and/or optical signal in the method may identify the polynucleotide as coming from a particular source.

[0218] For (iii), the sequence of the polynucleotide can be determined as described previously. Suitable sequencing methods, particularly those using electrical measurements, are described in Stoddart D et al., Proc Natl Acad Sci, 12; 106(19):7702-7, Lieberman K R et al, J Am Chem Soc. 2010; 132(50):17961-72, and International Application WO 2000/28312.

[0219] For (iv), the secondary structure may be measured in a variety of ways. For instance, if the method involves an electrical measurement, the secondary structure may be measured using a change in dwell time or a change in current flowing through the pore. This allows regions of single-stranded and double-stranded polynucleotide to be distinguished.

[0220] For (v), the presence or absence of any modification may be measured. The method preferably comprises determining whether or not the polynucleotide is modified by methylation, by oxidation, by damage, with one or more proteins or with one or more labels, tags or spacers. Specific modifications will result in specific interactions with the pore which can be measured using the methods described below. For instance, methylcyotsine may be distinguished from cytosine on the basis of the current flowing through the pore during its interaction with each nucleotide.

[0221] The target polynucleotide is contacted with a pore of the invention. The pore is typically present in a membrane. Suitable membranes are discussed below. The method may be carried out using any apparatus that is suitable for investigating a membrane/pore system in which a pore is present in a membrane. The method may be carried out using any apparatus that is suitable for transmembrane pore sensing. For example, the apparatus comprises a chamber comprising an aqueous solution and a barrier that separates the chamber into two sections. The barrier typically has an aperture in which the membrane containing the pore is formed. Alternatively the barrier forms the membrane in which the pore is present.

[0222] The method may be carried out using the apparatus described in International Application No. PCT/GB08/000562 (WO 2008/102120).

[0223] A variety of different types of measurements may be made. This includes without limitation: electrical measurements and optical measurements. Possible electrical measurements include: current measurements, impedance measurements, tunnelling measurements (Ivanov A P et al., Nano Lett. 2011 Jan. 12; 11(1):279-85), and FET measurements (International Application WO 2005/124888). Optical measurements may be combined with electrical measurements (Soni G V et al., Rev Sci Instrum. 2010 January;

81(1):014301). The measurement may be a transmembrane current measurement such as measurement of ionic current flowing through the pore.

[0224] Electrical measurements may be made using standard single channel recording equipment as describe in Stoddart D et al., Proc Natl Acad Sci, 12; 106(19):7702-7, Lieberman K R et al, J Am Chem Soc. 2010; 132(50):17961-72, and International Application WO 2000/28312. Alternatively, electrical measurements may be made using a multichannel system, for example as described in International Application WO 2009/077734 and International Application WO 2011/067559.

[0225] The method is preferably carried out with a potential applied across the membrane. The applied potential may be a voltage potential. Alternatively, the applied potential may be a chemical potential. An example of this is using a salt gradient across a membrane, such as an amphiphilic layer. A salt gradient is disclosed in Holden et al., J Am Chem Soc. 2007 Jul. 11; 129(27):8650-5. In some instances, the current passing through the pore as a polynucleotide moves with respect to the pore is used to estimate or determine the sequence of the polynucleotide. This is strand sequencing.

[0226] The method may involve measuring the current passing through the pore as the polynucleotide moves with respect to the pore. Therefore the apparatus used in the method may also comprise an electrical circuit capable of applying a potential and measuring an electrical signal across the membrane and pore. The methods may be carried out using a patch clamp or a voltage clamp. The methods preferably involve the use of a voltage clamp.

[0227] The method of the invention may involve the measuring of a current passing through the pore as the polynucleotide moves with respect to the pore. Suitable conditions for measuring ionic currents through transmembrane protein pores are known in the art and disclosed in the Example. The method is typically carried out with a voltage applied across the membrane and pore. The voltage used is typically from +5 V to −5 V, such as from +4 V to −4 V, +3 V to −3 V or +2 V to −2 V. The voltage used is typically from −600 mV to +600 mV or −400 mV to +400 mV. The voltage used is preferably in a range having a lower limit selected from −400 mV, −300 mV, −200 mV, −150 mV, −100 mV, −50 mV, −20 mV and 0 mV and an upper limit independently selected from +10 mV, +20 mV, +50 mV, +100 mV, +150 mV, +200 mV, +300 mV and +400 mV. The voltage used is more preferably in the range 100 mV to 240 mV and most preferably in the range of 120 mV to 220 mV. It is possible to increase discrimination between different nucleotides by a pore by using an increased applied potential.

[0228] The method is typically carried out in the presence of any charge carriers, such as metal salts, for example alkali metal salt, halide salts, for example chloride salts, such as alkali metal chloride salt. Charge carriers may include ionic liquids or organic salts, for example tetramethyl ammonium chloride, trimethylphenyl ammonium chloride, phenyltrimethyl ammonium chloride, or 1-ethyl-3-methyl imidazolium chloride. In the exemplary apparatus discussed above, the salt is present in the aqueous solution in the chamber. Potassium chloride (KCl), sodium chloride (NaCl), caesium chloride (CsCl) or a mixture of potassium ferrocyanide and potassium ferricyanide is typically used. KCl, NaCl and a mixture of potassium ferrocyanide and potassium ferricyanide are preferred. The charge carriers may be asymmetric across the membrane. For instance, the type and/or concentration of the charge carriers may be different on each side of the membrane.

[0229] The salt concentration may be at saturation. The salt concentration may be 3 M or lower and is typically from 0.1 to 2.5 M, from 0.3 to 1.9 M, from 0.5 to 1.8 M, from 0.7 to 1.7 M, from 0.9 to 1.6 M or from 1 M to 1.4 M. The salt concentration is preferably from 150 mM to 1 M. The method is preferably carried out using a salt concentration of at least 0.3 M, such as at least 0.4 M, at least 0.5 M, at least 0.6 M, at least 0.8 M, at least 1.0 M, at least 1.5 M, at least 2.0 M, at least 2.5 M or at least 3.0 M. High salt concentrations provide a high signal to noise ratio and allow for currents indicative of the presence of a nucleotide to be identified against the background of normal current fluctuations.

[0230] The method is typically carried out in the presence of a buffer. In the exemplary apparatus discussed above, the buffer is present in the aqueous solution in the chamber. Any buffer may be used in the method of the invention. Typically, the buffer is phosphate buffer. Other suitable buffers are HEPES and Tris-HCl buffer. The methods are typically carried out at a pH of from 4.0 to 12.0, from 4.5 to 10.0, from 5.0 to 9.0, from 5.5 to 8.8, from 6.0 to 8.7 or from 7.0 to 8.8 or 7.5 to 8.5. The pH used is preferably about 7.5.

[0231] The method may be carried out at from 0° C. to 100° C., from 15° C. to 95° C., from 16° C. to 90° C., from 17° C. to 85° C., from 18° C. to 80° C., 19° C. to 70° C., or from 20° C. to 60° C. The methods are typically carried out at room temperature. The methods are optionally carried out at a temperature that supports enzyme function, such as about 37° C.

Polynucleotide Binding Protein

[0232] Step (a) preferably comprises contacting the polynucleotide with a polynucleotide binding protein such that the protein controls the movement of the polynucleotide through the pore.

[0233] More preferably, the method comprises (a) contacting the polynucleotide with the pore of the invention and a polynucleotide binding protein such that the protein controls the movement of the polynucleotide through the pore and (b) taking one or more measurements as the polynucleotide moves with respect to the pore, wherein the measurements are indicative of one or more characteristics of the polynucleotide, and thereby characterising the polynucleotide.

[0234] More preferably, the method comprises (a) contacting the polynucleotide with the pore of the invention and a polynucleotide binding protein such that the protein controls the movement of the polynucleotide through the pore and (b) measuring the current through the pore as the polynucleotide moves with respect to the pore, wherein the current is indicative of one or more characteristics of the polynucleotide, and thereby characterising the polynucleotide.

[0235] The polynucleotide binding protein may be any protein that is capable of binding to the polynucleotide and controlling its movement through the pore. It is straightforward in the art to determine whether or not a protein binds to a polynucleotide. The protein typically interacts with and modifies at least one property of the polynucleotide. The protein may modify the polynucleotide by cleaving it to

form individual nucleotides or shorter chains of nucleotides, such as di- or trinucleotides. The protein may modify the polynucleotide by orienting it or moving it to a specific position, i.e. controlling its movement.

[0236] The polynucleotide binding protein is preferably derived from a polynucleotide handling enzyme. A polynucleotide handling enzyme is a polypeptide that is capable of interacting with and modifying at least one property of a polynucleotide. The enzyme may modify the polynucleotide by cleaving it to form individual nucleotides or shorter chains of nucleotides, such as di- or trinucleotides. The enzyme may modify the polynucleotide by orienting it or moving it to a specific position. The polynucleotide handling enzyme does not need to display enzymatic activity as long as it is capable of binding the polynucleotide and controlling its movement through the pore. For instance, the enzyme may be modified to remove its enzymatic activity or may be used under conditions which prevent it from acting as an enzyme. Such conditions are discussed in more detail below.

[0237] The polynucleotide handling enzyme is preferably derived from a nucleolytic enzyme. The polynucleotide handling enzyme used in the construct of the enzyme is more preferably derived from a member of any of the Enzyme Classification (EC) groups 3.1.11, 3.1.13, 3.1.14, 3.1.15, 3.1.16, 3.1.21, 3.1.22, 3.1.25, 3.1.26, 3.1.27, 3.1.30 and 3.1.31. The enzyme may be any of those disclosed in International Application No. PCT/GB10/000133 (published as WO 2010/086603).

[0238] Preferred enzymes are polymerases, exonucleases, helicases and topoisomerases, such as gyrases. Suitable enzymes include, but are not limited to, exonuclease I from *E. coli* (SEQ ID NO: 11), exonuclease III enzyme from *E. coli* (SEQ ID NO: 13), RecJ from *T. thermophilus* (SEQ ID NO: 15) and bacteriophage lambda exonuclease (SEQ ID NO: 17) and variants thereof. Three subunits comprising the sequence shown in SEQ ID NO: 15 or a variant thereof interact to form a trimer exonuclease. The enzyme is preferably Phi29 DNA polymerase (SEQ ID NO: 9) or a variant thereof. The topoisomerase is preferably a member of any of the Moiety Classification (EC) groups 5.99.1.2 and 5.99.1.3.

[0239] The enzyme is most preferably derived from a helicase, such as Hel308 Mbu (SEQ ID NO: 18), Hel308 Csy (SEQ ID NO: 19), Hel308 Mhu (SEQ ID NO: 21), TraI Eco (SEQ ID NO: 22), XPD Mbu (SEQ ID NO: 23) or a variant thereof. Any helicase may be used in the invention. The helicase may be or be derived from a Hel308 helicase, a RecD helicase, such as TraI helicase or a TrwC helicase, a XPD helicase or a Dda helicase. The helicase may be any of the helicases, modified helicases or helicase constructs disclosed in International Application Nos. PCT/GB2012/052579 (published as WO 2013/057495); PCT/GB2012/053274 (published as WO 2013/098562); PCT/GB2012/053273 (published as WO2013098561); PCT/GB2013/051925 (published as WO 2014/013260); PCT/GB2013/051924 (published as WO 2014/013259) and PCT/GB2013/051928 (published as WO 2014/013262); and in PCT/GB2014/052736 (published as WO/2015/055981).

[0240] The helicase preferably comprises the sequence shown in SEQ ID NO: 25 (Trwc Cba) or as variant thereof, the sequence shown in SEQ ID NO: 18 (Hel308 Mbu) or a variant thereof or the sequence shown in SEQ ID NO: 24 (Dda) or a variant thereof. Variants may differ from the native sequences in any of the ways discussed below for transmembrane pores. A preferred variant of SEQ ID NO: 8 comprises E94C/A360C and then (ΔM1)G1G2 (i.e. deletion of M1 and then addition G1 and G2).

[0241] Any number of helicases may be used in accordance with the invention. For instance, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more helicases may be used. In some embodiments, different numbers of helicases may be used.

[0242] The method of the invention preferably comprises contacting the polynucleotide with two or more helicases. The two or more helicases are typically the same helicase. The two or more helicases may be different helicases.

[0243] The two or more helicases may be any combination of the helicases mentioned above. The two or more helicases may be two or more Dda helicases. The two or more helicases may be one or more Dda helicases and one or more TrwC helicases. The two or more helicases may be different variants of the same helicase.

[0244] The two or more helicases are preferably attached to one another. The two or more helicases are more preferably covalently attached to one another. The helicases may be attached in any order and using any method. Preferred helicase constructs for use in the invention are described in International Application Nos. PCT/GB2013/051925 (published as WO 2014/013260); PCT/GB2013/051924 (published as WO 2014/013259) and PCT/GB2013/051928 (published as WO 2014/013262); and in PCT/GB2014/052736 (published as WO/2015/055981).

[0245] A variant of SEQ ID NOs: 9, 11, 13, 15, 17, 18, 19, 20, 21, 22, 23, 24 or 25 is an enzyme that has an amino acid sequence which varies from that of SEQ ID NO: 9, 11, 13, 15, 17, 18, 19, 20, 21, 22, 23, 24 or 25 and which retains polynucleotide binding ability. This can be measured using any method known in the art. For instance, the variant can be contacted with a polynucleotide and its ability to bind to and move along the polynucleotide can be measured. The variant may include modifications that facilitate binding of the polynucleotide and/or facilitate its activity at high salt concentrations and/or room temperature. Variants may be modified such that they bind polynucleotides (i.e. retain polynucleotide binding ability) but do not function as a helicase (i.e. do not move along polynucleotides when provided with all the necessary components to facilitate movement, e.g. ATP and Mg^{2+}). Such modifications are known in the art. For instance, modification of the Mg^{2+} binding domain in helicases typically results in variants which do not function as helicases. These types of variants may act as molecular brakes (see below).

[0246] Over the entire length of the amino acid sequence of SEQ ID NO: 9, 11, 13, 15, 17, 18, 19, 20, 21, 22, 23, 24 or 25, a variant will preferably be at least 50% homologous to that sequence based on amino acid identity. More preferably, the variant polypeptide may be at least 55%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90% and more preferably at least 95%, 97% or 99% homologous based on amino acid identity to the amino acid sequence of SEQ ID NO: 9, 11, 13, 15, 17, 18, 19, 20, 21, 22, 23, 24 or 25 over the entire sequence. There may be at least 80%, for example at least 85%, 90% or 95%, amino acid identity over a stretch of 200 or more, for example 230, 250, 270, 280, 300, 400, 500, 600, 700, 800, 900 or 1000 or more, contiguous amino acids ("hard homology"). Homology is determined as described above. The variant may differ from the wild-type sequence in any of the ways discussed above with reference to SEQ ID NO: 2 and

4 above. The enzyme may be covalently attached to the pore. Any method may be used to covalently attach the enzyme to the pore.

[0247] A preferred molecular brake is TrwC Cba-Q594A (SEQ ID NO: 25 with the mutation Q594A). This variant does not function as a helicase (i.e. binds polynucleotides but does not move along them when provided with all the necessary components to facilitate movement, e.g. ATP and Mg^{2+}).

[0248] In strand sequencing, the polynucleotide is translocated through the pore either with or against an applied potential. Exonucleases that act progressively or processively on double stranded polynucleotides can be used on the cis side of the pore to feed the remaining single strand through under an applied potential or the trans side under a reverse potential. Likewise, a helicase that unwinds the double stranded DNA can also be used in a similar manner. A polymerase may also be used. There are also possibilities for sequencing applications that require strand translocation against an applied potential, but the DNA must be first "caught" by the enzyme under a reverse or no potential. With the potential then switched back following binding the strand will pass cis to trans through the pore and be held in an extended conformation by the current flow. The single strand DNA exonucleases or single strand DNA dependent polymerases can act as molecular motors to pull the recently translocated single strand back through the pore in a controlled stepwise manner, trans to cis, against the applied potential.

[0249] Any helicase may be used in the method. Helicases may work in two modes with respect to the pore. First, the method is preferably carried out using a helicase such that it moves the polynucleotide through the pore with the field resulting from the applied voltage. In this mode the 5' end of the polynucleotide is first captured in the pore, and the helicase moves the polynucleotide into the pore such that it is passed through the pore with the field until it finally translocates through to the trans side of the membrane. Alternatively, the method is preferably carried out such that a helicase moves the polynucleotide through the pore against the field resulting from the applied voltage. In this mode the 3' end of the polynucleotide is first captured in the pore, and the helicase moves the polynucleotide through the pore such that it is pulled out of the pore against the applied field until finally ejected back to the cis side of the membrane.

[0250] The method may also be carried out in the opposite direction. The 3' end of the polynucleotide may be first captured in the pore and the helicase may move the polynucleotide into the pore such that it is passed through the pore with the field until it finally translocates through to the trans side of the membrane.

[0251] When the helicase is not provided with the necessary components to facilitate movement or is modified to hinder or prevent its movement, it can bind to the polynucleotide and act as a brake slowing the movement of the polynucleotide when it is pulled into the pore by the applied field. In the inactive mode, it does not matter whether the polynucleotide is captured either 3' or 5' down, it is the applied field which pulls the polynucleotide into the pore towards the trans side with the enzyme acting as a brake. When in the inactive mode, the movement control of the polynucleotide by the helicase can be described in a number of ways including ratcheting, sliding and braking. Helicase variants which lack helicase activity can also be used in this way.

[0252] The polynucleotide may be contacted with the polynucleotide binding protein and the pore in any order. It is preferred that, when the polynucleotide is contacted with the polynucleotide binding protein, such as a helicase, and the pore, the polynucleotide firstly forms a complex with the protein. When the voltage is applied across the pore, the polynucleotide/protein complex then forms a complex with the pore and controls the movement of the polynucleotide through the pore.

[0253] Any steps in the method using a polynucleotide binding protein are typically carried out in the presence of free nucleotides or free nucleotide analogues and an enzyme cofactor that facilitates the action of the polynucleotide binding protein. The free nucleotides may be one or more of any of the individual nucleotides discussed above. The free nucleotides include, but are not limited to, adenosine monophosphate (AMP), adenosine diphosphate (ADP), adenosine triphosphate (ATP), guanosine monophosphate (GMP), guanosine diphosphate (GDP), guanosine triphosphate (GTP), thymidine monophosphate (TMP), thymidine diphosphate (TDP), thymidine triphosphate (TTP), uridine monophosphate (UMP), uridine diphosphate (UDP), uridine triphosphate (UTP), cytidine monophosphate (CMP), cytidine diphosphate (CDP), cytidine triphosphate (CTP), cyclic adenosine monophosphate (cAMP), cyclic guanosine monophosphate (cGMP), deoxyadenosine monophosphate (dAMP), deoxyadenosine diphosphate (dADP), deoxyadenosine triphosphate (dATP), deoxyguanosine monophosphate (dGMP), deoxyguanosine diphosphate (dGDP), deoxyguanosine triphosphate (dGTP), deoxythymidine monophosphate (dTMP), deoxythymidine diphosphate (dTDP), deoxythymidine triphosphate (dTTP), deoxyuridine monophosphate (dUMP), deoxyuridine diphosphate (dUDP), deoxyuridine triphosphate (dUTP), deoxycytidine monophosphate (dCMP), deoxycytidine diphosphate (dCDP) and deoxycytidine triphosphate (dCTP). The free nucleotides are preferably selected from AMP, TMP, GMP, CMP, UMP, dAMP, dTMP, dGMP or dCMP. The free nucleotides are preferably adenosine triphosphate (ATP). The enzyme cofactor is a factor that allows the construct to function. The enzyme cofactor is preferably a divalent metal cation. The divalent metal cation is preferably Mg^{2+}, Mn^{2+}, Ca^{2+} or Co^{2+}. The enzyme cofactor is most preferably Mg^{2+}.

Helicase(s) and Molecular Brake(s)

[0254] In a preferred embodiment, the method comprises:

[0255] (a) providing the polynucleotide with one or more helicases and one or more molecular brakes attached to the polynucleotide;

[0256] (b) contacting the polynucleotide with a pore of the invention and applying a potential across the pore such that the one or more helicases and the one or more molecular brakes are brought together and both control the movement of the polynucleotide through the pore;

[0257] (c) taking one or more measurements as the polynucleotide moves with respect to the pore wherein the measurements are indicative of one or more characteristics of the polynucleotide and thereby characterising the polynucleotide. This type of method is discussed in detail in International Application No. PCT/GB2014/052737.

[0258] The one or more helicases may be any of those discussed above. The one or more molecular brakes may be any compound or molecule which binds to the polynucleotide and slow the movement of the polynucleotide through the pore. The one or more molecular brakes preferably comprise one or more compounds which bind to the polynucleotide. The one or more compounds are preferably one or more macrocycles. Suitable macrocycles include, but are not limited to, cyclodextrins, calixarenes, cyclic peptides, crown ethers, cucurbiturils, pillararenes, derivatives thereof or a combination thereof. The cyclodextrin or derivative thereof may be any of those disclosed in Eliseev, A. V., and Schneider, H-J. (1994) *J. Am. Chem. Soc.* 116, 6081-6088. The agent is more preferably heptakis-6-amino-β-cyclodextrin (am_7-βCD), 6-monodeoxy-6-monoamino-β-cyclodextrin (am_1-βCD) or heptakis-(6-deoxy-6-guanidino)-cyclodextrin (gu_7-βCD).

[0259] The one or more molecular brakes are preferably not one or more single stranded binding proteins (SSB). The one or more molecular brakes are more preferably not a single-stranded binding protein (SSB) comprising a carboxy-terminal (C-terminal) region which does not have a net negative charge or (ii) a modified SSB comprising one or more modifications in its C-terminal region which decreases the net negative charge of the C-terminal region. The one or more molecular brakes are most preferably not any of the SSBs disclosed in International Application No. PCT/GB2013/051924 (published as WO 2014/013259).

[0260] The one or more molecular brakes are preferably one or more polynucleotide binding proteins. The polynucleotide binding protein may be any protein that is capable of binding to the polynucleotide and controlling its movement through the pore. It is straightforward in the art to determine whether or not a protein binds to a polynucleotide. The protein typically interacts with and modifies at least one property of the polynucleotide. The protein may modify the polynucleotide by cleaving it to form individual nucleotides or shorter chains of nucleotides, such as di- or trinucleotides. The moiety may modify the polynucleotide by orienting it or moving it to a specific position, i.e. controlling its movement.

[0261] The polynucleotide binding protein is preferably derived from a polynucleotide handling enzyme. The one or more molecular brakes may be derived from any of the polynucleotide handling enzymes discussed above. Modified versions of Phi29 polymerase (SEQ ID NO: 8) which act as molecular brakes are disclosed in U.S. Pat. No. 5,576,204. The one or more molecular brakes are preferably derived from a helicase.

[0262] Any number of molecular brakes derived from a helicase may be used. For instance, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more helicases may be used as molecular brakes. If two or more helicases are be used as molecular brakes, the two or more helicases are typically the same helicase. The two or more helicases may be different helicases.

[0263] The two or more helicases may be any combination of the helicases mentioned above. The two or more helicases may be two or more Dda helicases. The two or more helicases may be one or more Dda helicases and one or more TrwC helicases. The two or more helicases may be different variants of the same helicase.

[0264] The two or more helicases are preferably attached to one another. The two or more helicases are more preferably covalently attached to one another. The helicases may be attached in any order and using any method. The one or more molecular brakes derived from helicases are preferably modified to reduce the size of an opening in the polynucleotide binding domain through which in at least one conformational state the polynucleotide can unbind from the helicase. This is disclosed in WO 2014/013260.

[0265] Preferred helicase constructs for use in the invention are described in International Application Nos. PCT/GB2013/051925 (published as WO 2014/013260); PCT/GB2013/051924 (published as WO 2014/013259) and PCT/GB2013/051928 (published as WO 2014/013262); and in PCT/GB2014/052736 (published as WO/2015/055981).

Spacers

[0266] The one or more helicases may be stalled at one or more spacers as discussed in International Application No. PCT/GB2014/050175. Any configuration of one or more helicases and one or more spacers disclosed in the International Application may be used in this invention.

Membrane

[0267] The pore of the invention may be present in a membrane. In the method of the invention, the polynucleotide is typically contacted with the pore of the invention in a membrane. Any membrane may be used in accordance with the invention. Suitable membranes are well-known in the art. The membrane is preferably an amphiphilic layer. An amphiphilic layer is a layer formed from amphiphilic molecules, such as phospholipids, which have both hydrophilic and lipophilic properties. The amphiphilic molecules may be synthetic or naturally occurring. Non-naturally occurring amphiphiles and amphiphiles which form a monolayer are known in the art and include, for example, block copolymers (Gonzalez-Perez et al., Langmuir, 2009, 25, 10447-10450). Block copolymers are polymeric materials in which two or more monomer sub-units that are polymerized together to create a single polymer chain. Block copolymers typically have properties that are contributed by each monomer sub-unit. However, a block copolymer may have unique properties that polymers formed from the individual sub-units do not possess. Block copolymers can be engineered such that one of the monomer sub-units is hydrophobic (i.e. lipophilic), whilst the other sub-unit(s) are hydrophilic whilst in aqueous media. In this case, the block copolymer may possess amphiphilic properties and may form a structure that mimics a biological membrane. The block copolymer may be a diblock (consisting of two monomer sub-units), but may also be constructed from more than two monomer sub-units to form more complex arrangements that behave as amphipiles. The copolymer may be a triblock, tetrablock or pentablock copolymer. The membrane is preferably a triblock copolymer membrane.

[0268] Archaebacterial bipolar tetraether lipids are naturally occurring lipids that are constructed such that the lipid forms a monolayer membrane. These lipids are generally found in extremophiles that survive in harsh biological environments, thermophiles, halophiles and acidophiles. Their stability is believed to derive from the fused nature of the final bilayer. It is straightforward to construct block copolymer materials that mimic these biological entities by creating a triblock polymer that has the general motif hydrophilic-hydrophobic-hydrophilic. This material may form monomeric membranes that behave similarly to lipid

bilayers and encompass a range of phase behaviours from vesicles through to laminar membranes. Membranes formed from these triblock copolymers hold several advantages over biological lipid membranes. Because the triblock copolymer is synthesized, the exact construction can be carefully controlled to provide the correct chain lengths and properties required to form membranes and to interact with pores and other proteins.

[0269] Block copolymers may also be constructed from sub-units that are not classed as lipid sub-materials; for example a hydrophobic polymer may be made from siloxane or other non-hydrocarbon based monomers. The hydrophilic sub-section of block copolymer can also possess low protein binding properties, which allows the creation of a membrane that is highly resistant when exposed to raw biological samples. This head group unit may also be derived from non-classical lipid head-groups.

[0270] Triblock copolymer membranes also have increased mechanical and environmental stability compared with biological lipid membranes, for example a much higher operational temperature or pH range. The synthetic nature of the block copolymers provides a platform to customize polymer based membranes for a wide range of applications.

[0271] The membrane is most preferably one of the membranes disclosed in International Application No. PCT/GB2013/052766 or PCT/GB2013/052767.

[0272] The amphiphilic molecules may be chemically-modified or functionalised to facilitate coupling of the polynucleotide.

[0273] The amphiphilic layer may be a monolayer or a bilayer. The amphiphilic layer is typically planar. The amphiphilic layer may be curved. The amphiphilic layer may be supported.

[0274] Amphiphilic membranes are typically naturally mobile, essentially acting as two dimensional fluids with lipid diffusion rates of approximately 10^8 cm s-1. This means that the pore and coupled polynucleotide can typically move within an amphiphilic membrane.

[0275] The membrane may be a lipid bilayer. Lipid bilayers are models of cell membranes and serve as excellent platforms for a range of experimental studies. For example, lipid bilayers can be used for in vitro investigation of membrane proteins by single-channel recording. Alternatively, lipid bilayers can be used as biosensors to detect the presence of a range of substances. The lipid bilayer may be any lipid bilayer. Suitable lipid bilayers include, but are not limited to, a planar lipid bilayer, a supported bilayer or a liposome. The lipid bilayer is preferably a planar lipid bilayer. Suitable lipid bilayers are disclosed in International Application No. PCT/GB08/000563 (published as WO 2008/102121), International Application No. PCT/GB08/004127 (published as WO 2009/077734) and International Application No. PCT/GB2006/001057 (published as WO 2006/100484).

Coupling

[0276] The polynucleotide is preferably coupled to the membrane comprising the pore of the invention. The method may comprise coupling the polynucleotide to the membrane comprising the pore of the invention. The polynucleotide is preferably coupled to the membrane using one or more anchors. The polynucleotide may be coupled to the membrane using any known method.

[0277] Each anchor comprises a group which couples (or binds) to the polynucleotide and a group which couples (or binds) to the membrane. Each anchor may covalently couple (or bind) to the polynucleotide and/or the membrane. If a Y adaptor and/or a hairpin loop adaptors are used, the polynucleotide is preferably coupled to the membrane using the adaptor(s).

[0278] The polynucleotide may be coupled to the membrane using any number of anchors, such as 2, 3, 4 or more anchors. For instance, a polynucleotide may be coupled to the membrane using two anchors each of which separately couples (or binds) to both the polynucleotide and membrane.

[0279] The one or more anchors may comprise the one or more helicases and/or the one or more molecular brakes.

[0280] If the membrane is an amphiphilic layer, such as a copolymer membrane or a lipid bilayer, the one or more anchors preferably comprise a polypeptide anchor present in the membrane and/or a hydrophobic anchor present in the membrane. The hydrophobic anchor is preferably a lipid, fatty acid, sterol, carbon nanotube, polypeptide, protein or amino acid, for example cholesterol, palmitate or tocopherol. In preferred embodiments, the one or more anchors are not the pore.

[0281] The components of the membrane, such as the amphiphilic molecules, copolymer or lipids, may be chemically-modified or functionalised to form the one or more anchors. Examples of suitable chemical modifications and suitable ways of functionalising the components of the membrane are discussed in more detail below. Any proportion of the membrane components may be functionalized, for example at least 0.01%, at least 0.1%, at least 1%, at least 10%, at least 25%, at least 50% or 100%.

[0282] The polynucleotide may be coupled directly to the membrane. The one or more anchors used to couple the polynucleotide to the membrane preferably comprise a linker. The one or more anchors may comprise one or more, such as 2, 3, 4 or more, linkers. One linker may be used couple more than one, such as 2, 3, 4 or more, polynucleotides to the membrane.

[0283] Preferred linkers include, but are not limited to, polymers, such as polynucleotides, polyethylene glycols (PEGs), polysaccharides and polypeptides. These linkers may be linear, branched or circular. For instance, the linker may be a circular polynucleotide. The polynucleotide may hybridise to a complementary sequence on the circular polynucleotide linker.

[0284] The coupling may be permanent or stable. In other words, the coupling may be such that the polynucleotide remains coupled to the membrane when interacting with the pore.

[0285] The coupling may be transient. In other words, the coupling may be such that the polynucleotide may decouple from the membrane when interacting with the pore.

[0286] Coupling of polynucleotides to a linker or to a functionalised membrane can also be achieved by a number of other means provided that a complementary reactive group or an anchoring group can be added to the polynucleotide. The addition of reactive groups to either end of a polynucleotide has been reported previously. The one or more anchors preferably couple the polynucleotide to the membrane via hybridisation. Hybridisation in the one or more anchors allows coupling in a transient manner as discussed above. The one or more anchors may comprise a single stranded or double stranded polynucleotide. One part

of the anchor may be ligated to a single stranded or double stranded polynucleotide. Ligation of short pieces of ssDNA have been reported using T4 RNA ligase I (Troutt, A. B., M. G. McHeyzer-Williams, et al. (1992). "Ligation-anchored PCR: a simple amplification technique with single-sided specificity." Proc Natl Acad Sci USA 89(20): 9823-5). Alternatively, either a single stranded or double stranded polynucleotide can be ligated to a double stranded polynucleotide and then the two strands separated by thermal or chemical denaturation. If the polynucleotide is a synthetic strand, the one or more anchors can be incorporated during the chemical synthesis of the polynucleotide. For instance, the polynucleotide can be synthesised using a primer having a reactive group attached to it.

[0287] Ideally, the polynucleotide is coupled to the membrane without having to functionalise the polynucleotide. This can be achieved by coupling the one or more anchors, such as a polynucleotide binding protein or a chemical group, to the membrane and allowing the one or more anchors to interact with the polynucleotide or by functionalizing the membrane. The one or more anchors may be coupled to the membrane by any of the methods described herein. In particular, the one or more anchors may comprise one or more linkers, such as maleimide functionalised linkers.

[0288] In this embodiment, the polynucleotide is typically RNA, DNA, PNA, TNA or LNA and may be double or single stranded. This embodiment is particularly suited to genomic DNA polynucleotides.

[0289] Where the one or more anchors comprise a protein, they may be able to anchor directly into the membrane without further functonalisation, for example if it already has an external hydrophobic region which is compatible with the membrane. Examples of such proteins include, but are not limited to, transmembrane proteins, intramembrane proteins and membrane proteins.

[0290] According to a preferred embodiment, the one or more anchors may be used to couple a polynucleotide to the membrane when the polynucleotide is attached to a leader sequence which preferentially threads into the pore. Leader sequences are discussed in more detail below. Preferably, the polynucleotide is attached (such as ligated) to a leader sequence which preferentially threads into the pore. Such a leader sequence may comprise a homopolymeric polynucleotide or an abasic region. The leader sequence is typically designed to hybridise to the one or more anchors either directly or via one or more intermediate polynucleotides (or splints). In such instances, the one or more anchors typically comprise a polynucleotide sequence which is complementary to a sequence in the leader sequence or a sequence in the one or more intermediate polynucleotides (or splints). In such instances, the one or more splints typically comprise a polynucleotide sequence which is complementary to a sequence in the leader sequence.

Double Stranded Polynucleotide

[0291] The polynucleotide may be double stranded. If the polynucleotide is double stranded, the method preferably further comprises before the contacting step ligating a hairpin adaptor to one end of the polynucleotide. The two strands of the polynucleotide may then be separated as or before the polynucleotide is contacted with the pore in accordance with the invention. The two strands may be separated as the polynucleotide movement through the pore is controlled by a polynucleotide binding protein, such as a helicase, or molecular brake. This is described in International Application No. PCT/GB2012/051786 (published as WO 2013/014451). Linking and interrogating both strands on a double stranded construct in this way increases the efficiency and accuracy of characterization.

Round the Corner Sequencing

[0292] In a preferred embodiment, a target double stranded polynucleotide is provided with a hairpin loop adaptor at one end and the method comprises contacting the polynucleotide with the pore of the invention such that both strands of the polynucleotide move through the pore and taking one or more measurements as the both strands of the polynucleotide move with respect to the pore wherein the measurements are indicative of one or more characteristics of the strands of the polynucleotide and thereby characterising the target double stranded polynucleotide. Any of the embodiments discussed above equally apply to this embodiment.

Leader Sequence

[0293] Before the contacting step, the method preferably comprises attaching to the polynucleotide a leader sequence which preferentially threads into the pore. The leader sequence facilitates the method of the invention. The leader sequence is designed to preferentially thread into the pore of the invention and thereby facilitate the movement of polynucleotide through the pore. The leader sequence can also be used to link the polynucleotide to the one or more anchors as discussed above.

Double Coupling

[0294] The method of the invention may involve double coupling of a double stranded polynucleotide. In a preferred embodiment, the method of the invention comprises:

[0295] (a) providing the double stranded polynucleotide with a Y adaptor at one end and a hairpin loop adaptor at the other end, wherein the Y adaptor comprises one or more first anchors for coupling the polynucleotide to the membrane, wherein the hairpin loop adaptor comprises one or more second anchors for coupling the polynucleotide to the membrane and wherein the strength of coupling of the hairpin loop adaptor to the membrane is greater than the strength of coupling of the Y adaptor to the membrane;

[0296] (b) contacting the polynucleotide provided in step (a) with the pore the invention such that the polynucleotide moves through the pore; and

[0297] (c) taking one or more measurements as the polynucleotide moves with respect to the pore, wherein the measurements are indicative of one or more characteristics of the polynucleotide, and thereby characterising the target polynucleotide. This type of method is discussed in detail in International Application No. PCT/GB2015/050991.

Adding Hairpin Loops and Leader Sequences

[0298] Before provision, a double stranded polynucleotide may be contacted with a MuA transposase and a population of double stranded MuA substrates, wherein a proportion of the substrates in the population are Y adaptors comprising the leader sequence and wherein a proportion of the substrates in the population are hairpin loop adaptors. The transposase fragments the double stranded polynucleotide

analyte and ligates MuA substrates to one or both ends of the fragments. This produces a plurality of modified double stranded polynucleotides comprising the leader sequence at one end and the hairpin loop at the other. The modified double stranded polynucleotides may then be investigated using the method of the invention. These MuA based methods are disclosed in International Application No. PCT/GB2014/052505 (published as WO/2015/022544). They are also discussed in detail in International Application No. PCT/GB2015/050991.

[0299] One or more helicases may be attached to the MuA substrate Y adaptors before they are contacted with the double stranded polynucleotide and MuA transposase. Alternatively, one or more helicases may be attached to the MuA substrate Y adaptors before they are contacted with the double stranded polynucleotide and MuA transposase.

[0300] One or more molecular brakes may be attached to the MuA substrate hairpin loop adaptors before they are contacted with the double stranded polynucleotide and MuA transposase. Alternatively, one or more molecular brakes may be attached to the MuA substrate hairpin loop adaptors before they are contacted with the double stranded polynucleotide and MuA transposase.

Uncoupling

[0301] The method of the invention may involve characterising multiple target polynucleotides and uncoupling of the at least the first target polynucleotide.

[0302] In a preferred embodiment, the invention involves characterising two or more target polynucleotides. The method comprises:

[0303] (a) providing a first polynucleotide in a first sample;

[0304] (b) providing a second polynucleotide in a second sample;

[0305] (c) coupling the first polynucleotide in the first sample to a membrane using one or more anchors;

[0306] (d) contacting the first polynucleotide with the pore of the invention such that the polynucleotide moves through the pore;

[0307] (e) taking one or more measurements as the first polynucleotide moves with respect to the pore wherein the measurements are indicative of one or more characteristics of the first polynucleotide and thereby characterising the first polynucleotide;

[0308] (f) uncoupling the first polynucleotide from the membrane;

[0309] (g) coupling the second polynucleotide in the second sample to the membrane using one or more anchors;

[0310] (h) contacting the second polynucleotide with the pore of the invention such that the second polynucleotide moves through the pore; and

[0311] (i) taking one or more measurements as the second polynucleotide moves with respect to the pore wherein the measurements are indicative of one or more characteristics of the second polynucleotide and thereby characterising the second polynucleotide.

[0312] This type of method is discussed in detail in International Application No. PCT/GB2015/050992.

[0313] If one or more anchors comprise a hydrophobic anchor, such as cholesterol, the agent is preferably a cyclodextrin or a derivative thereof or a lipid. The cyclodextrin or derivative thereof may be any of those disclosed in Eliseev, A. V., and Schneider, H-J. (1994) *J. Am. Chem. Soc.* 116, 6081-6088. The agent is more preferably heptakis-6-amino-β-cyclodextrin (am_7-βCD), 6-monodeoxy-6-monoamino-β-cyclodextrin (am_1-βCD) or heptakis-(6-deoxy-6-guanidino)-cyclodextrin (gu_7-βCD). Any of the lipids disclosed herein may be used.

Modified Polynucleotides

[0314] Before characterisation, a target polynucleotide may be modified by contacting the polynucleotide with a polymerase and a population of free nucleotides under conditions in which the polymerase forms a modified polynucleotide using the target polynucleotide as a template, wherein the polymerase replaces one or more of the nucleotide species in the target polynucleotide with a different nucleotide species when forming the modified polynucleotide. The modified polynucleotide may then be provided with one or more helicases attached to the polynucleotide and one or more molecular brakes attached to the polynucleotide. This type of modification is described in International Application No. PCT/GB2015/050483. Any of the polymerases discussed above may be used. The polymerase is preferably Klenow or 90 North.

[0315] The template polynucleotide is contacted with the polymerase under conditions in which the polymerase forms a modified polynucleotide using the template polynucleotide as a template. Such conditions are known in the art. For instance, the polynucleotide is typically contacted with the polymerase in commercially available polymerase buffer, such as buffer from New England Biolabs®. The temperature is preferably from 20 to 37° C. for Klenow or from 60 to 75° C. for 90 North. A primer or a 3' hairpin is typically used as the nucleation point for polymerase extension.

[0316] Characterisation, such as sequencing, of a polynucleotide using a transmembrane pore typically involves analyzing polymer units made up of k nucleotides where k is a positive integer (i.e. 'k-mers'). This is discussed in International Application No. PCT/GB2012/052343 (published as WO 2013/041878). While it is desirable to have clear separation between current measurements for different k-mers, it is common for some of these measurements to overlap. Especially with high numbers of polymer units in the k-mer, i.e. high values of k, it can become difficult to resolve the measurements produced by different k-mers, to the detriment of deriving information about the polynucleotide, for example an estimate of the underlying sequence of the polynucleotide.

[0317] By replacing one or more nucleotide species in the target polynucleotide with different nucleotide species in the modified polynucleotide, the modified polynucleotide contains k-mers which differ from those in the target polynucleotide. The different k-mers in the modified polynucleotide are capable of producing different current measurements from the k-mers in the target polynucleotide and so the modified polynucleotide provides different information from the target polynucleotide. The additional information from the modified polynucleotide can make it easier to characterise the target polynucleotide. In some instances, the modified polynucleotide itself may be easier to characterise. For instance, the modified polynucleotide may be designed to include k-mers with an increased separation or a clear separation between their current measurements or k-mers which have a decreased noise.

[0318] The polymerase preferably replaces two or more of the nucleotide species in the target polynucleotide with different nucleotide species when forming the modified polynucleotide. The polymerase may replace each of the two or more nucleotide species in the target polynucleotide with a distinct nucleotide species. The polymerase may replace each of the two or more nucleotide species in the target polynucleotide with the same nucleotide species.

[0319] If the target polynucleotide is DNA, the different nucleotide species in the modified typically comprises a nucleobase which differs from adenine, guanine, thymine, cytosine or methylcytosine and/or comprises a nucleoside which differs from deoxyadenosine, deoxyguanosine, thymidine, deoxycytidine or deoxymethylcytidine. If the target polynucleotide is RNA, the different nucleotide species in the modified polynucleotide typically comprises a nucleobase which differs from adenine, guanine, uracil, cytosine or methylcytosine and/or comprises a nucleoside which differs from adenosine, guanosine, uridine, cytidine or methylcytidine. The different nucleotide species may be any of the universal nucleotides discussed above.

[0320] The polymerase may replace the one or more nucleotide species with a different nucleotide species which comprises a chemical group or atom absent from the one or more nucleotide species. The chemical group may be a propynyl group, a thio group, an oxo group, a methyl group, a hydroxymethyl group, a formyl group, a carboxy group, a carbonyl group, a benzyl group, a propargyl group or a propargylamine group.

[0321] The polymerase may replace the one or more nucleotide species with a different nucleotide species which lacks a chemical group or atom present in the one or more nucleotide species. The polymerase may replace the one or more of the nucleotide species with a different nucleotide species having an altered electronegativity. The different nucleotide species having an altered electronegativity preferably comprises a halogen atom.

[0322] The method preferably further comprises selectively removing the nucleobases from the one or more different nucleotides species in the modified polynucleotide.

Other Characterisation Method

[0323] In another embodiment, a polynucleotide is characterised by detecting labelled species that are released as a polymerase incorporates nucleotides into the polynucleotide. The polymerase uses the polynucleotide as a template. Each labelled species is specific for each nucleotide. The polynucleotide is contacted with a pore of the invention, a polymerase and labelled nucleotides such that phosphate labelled species are sequentially released when nucleotides are added to the polynucleotide(s) by the polymerase, wherein the phosphate species contain a label specific for each nucleotide. The polymerase may be any of those discussed above. The phosphate labelled species are detected using the pore and thereby characterising the polynucleotide. This type of method is disclosed in European Application No. 13187149.3 (published as EP 2682460). Any of the embodiments discussed above equally apply to this method.

Kits

[0324] The present invention also provides a kit for characterising a target polynucleotide. The kit comprises a pore of the invention and the components of a membrane. The membrane is preferably formed from the components. The pore is preferably present in the membrane. The kit may comprise components of any of the membranes disclosed above, such as an amphiphilic layer or a triblock copolymer membrane.

[0325] The kit may further comprise a polynucleotide binding protein.

[0326] The kit may further comprise one or more anchors for coupling the polynucleotide to the membrane.

[0327] The kit is preferably for characterising a double stranded polynucleotide and preferably comprises a Y adaptor and a hairpin loop adaptor. The Y adaptor preferably has one or more helicases attached and the hairpin loop adaptor preferably has one or more molecular brakes attached. The Y adaptor preferably comprises one or more first anchors for coupling the polynucleotide to the membrane, the hairpin loop adaptor preferably comprises one or more second anchors for coupling the polynucleotide to the membrane and the strength of coupling of the hairpin loop adaptor to the membrane is preferably greater than the strength of coupling of the Y adaptor to the membrane.

[0328] The kit of the invention may additionally comprise one or more other reagents or instruments which enable any of the embodiments mentioned above to be carried out. Such reagents or instruments include one or more of the following: suitable buffer(s) (aqueous solutions), means to obtain a sample from a subject (such as a vessel or an instrument comprising a needle), means to amplify and/or express polynucleotides or voltage or patch clamp apparatus. Reagents may be present in the kit in a dry state such that a fluid sample resuspends the reagents. The kit may also, optionally, comprise instructions to enable the kit to be used in the method of the invention or details regarding for which organism the method may be used.

Apparatus

[0329] The invention also provides an apparatus for characterising a target polynucleotide. The apparatus comprises a plurality of pores of the invention and a plurality of membranes. The plurality of pores are preferably present in the plurality of membranes. The number of pores and membranes is preferably equal. Preferably, a single pore is present in each membrane.

[0330] The apparatus preferably further comprises instructions for carrying out the method of the invention. The apparatus may be any conventional apparatus for polynucleotide analysis, such as an array or a chip. Any of the embodiments discussed above with reference to the methods of the invention are equally applicable to the apparatus of the invention. The apparatus may further comprise any of the features present in the kit of the invention.

[0331] The apparatus is preferably set up to carry out the method of the invention.

[0332] The apparatus preferably comprises:

[0333] a sensor device that is capable of supporting the plurality of pores and membranes and being operable to perform polynucleotide characterisation using the pores and membranes; and

[0334] at least one port for delivery of the material for performing the characterisation.

[0335] Alternatively, the apparatus preferably comprises:

[0336] a sensor device that is capable of supporting the plurality of pores and membranes being operable to perform polynucleotide characterisation using the pores and membranes; and

[0337] at least one reservoir for holding material for performing the characterisation.

[0338] The apparatus more preferably comprises:

[0339] a sensor device that is capable of supporting the membrane and plurality of pores and membranes and being operable to perform polynucleotide characterising using the pores and membranes;

[0340] at least one reservoir for holding material for performing the characterising;

[0341] a fluidics system configured to controllably supply material from the at least one reservoir to the sensor device; and

[0342] one or more containers for receiving respective samples, the fluidics system being configured to supply the samples selectively from one or more containers to the sensor device.

[0343] The apparatus may be any of those described in International Application No. PCT/GB08/004127 (published as WO 2009/077734), PCT/GB10/000789 (published as WO 2010/122293), International Application No. PCT/GB10/002206 (published as WO 2011/067559) or International Application No. PCT/US99/25679 (published as WO 00/28312).

[0344] The following Example illustrates the invention.

Example 1

[0345] This Example describes the use of a helicase—T4 Dda—E94C/C109A/C136A/A360C (SEQ ID NO: 24 with mutations E94C/C109A/C136A/A360C) to control the movement of DNA construct X (shown in FIG. **1**) through a number of different MspA nanopores. All of the nanopores tested exhibited changes in current as the DNA translocated through the nanopore.

Materials and Methods

[0346] Prior to setting up the experiment, DNA construct X or Y (see FIGS. **1** and **2** respectively for diagram and sequences used in constructs X and Y, final concentration added to the nanopore system 0.1 nM) was pre-incubated at room temperature for five minutes with T4 Dda—E94C/C109A/C136A/A360C (final concentration added to the nanopore system 10 nM, SEQ ID NO: 24 with mutations E94C/A360C which was provided in buffer (253 mM KCl, 50 mM potassium phosphate, pH 8.0, 2 mM EDTA)). After five minutes, TMAD (100 μM final concentration added to the nanopore system) was added to the pre-mix and the mixture incubated for a further 5 minutes. Finally, MgCl2 (1 or 2 mM final concentration added to the nanopore system), ATP (2 mM final concentration added to the nanopore system) and KCl (500 mM final concentration added to the nanopore system) were added to the pre-mix.

[0347] Electrical measurements were acquired from single MspA nanopores (please see table 5 below for those tested) inserted in block co-polymer in buffer (500 mM KCl, 25 mM potassium phosphate pH 8.0). After achieving a single pore inserted in the block co-polymer, then buffer (1 mL, 500 mM KCl, 25 mM potassium phosphate) was flowed through the system to remove any excess MspA nanopores. The enzyme (T4 Dda—E94C/C109A/C136A/A360C, 10 nM final concentration), DNA construct X or Y (0.1 nM final concentration), fuel (MgCl2 1 or 2 mM final concentration, ATP 2 mM final concentration) pre-mix (150 μL total) was then flowed into the single nanopore experimental system and the experiment run with the following protocol −120 mV for 900 s, −180 mV for 2 seconds, 0 mV for 2 s and then return to 120 mV for 900 s (this flip protocol was repeated 8 times) and helicase-controlled DNA movement was monitored.

Results

[0348] The MspA nanopores which were tested are shown in the table 3 below. Helicase controlled DNA movement was observed for DNA constructs X and Y (FIGS. **1** and **2** respectively) using T4 Dda—E94C/C109A/C136A/A360C (see table 5 for which figures correspond to which nanopore mutants tested).

TABLE 3

Entry	Mutant MspA Nanopore (SEQ ID NO: 2 with the specified mutations)	DNA Construct	Final Mg^{2+} concentration (mM)	Figure Number
1	MspA - (G75S/G77S/L88N/D90N/D91N/D118R/Q126R/D134R/E139K)8	X	1	3
2	MspA - (G75S/G77S/L88N/D90N/D91N/A96D/D118R/Q126R/D134R/E139K)8	Y	2	4
3	MspA - (G75S/G77S/L88N/D90N/D91N/N102G/D118R/Q126R/D134R/E139K)8	X	1	5
4	MspA - (G75S/G77S/L88N/D90N/D91N/S103A/D118R/Q126R/D134R/E139K)8	X	1	6
5	MspA - (G75S/G77S/L88N/D90N/D91N/N108S/D118R/Q126R/D134R/E139K)8	X	1	7
6	MspA - (G75S/G77S/L88N/D90N/D91N/N108P/D118R/Q126R/D134R/E139K)8	X	1	8
7	MspA - (G75S/G77S/L88N/D90N/D9IN/A96D/N108P/D118R/Q126R/D134R/E139K)8	X	1	9
8	MspA - (G75S/G77S/L88N/D90N/D9IN/A96D/N108A/D118R/Q126R/D134R/E139K)8	X	1	10
9	MspA - (G75S/G77S/L88N/I89F/D90N/D91N/D118R/Q126R/D134R/E139K)8	X	1	11
10	MspA - (G75S/G77S/D90N/D91N/D118R/Q126R/D134R/E139K)8	X	1	12
11	MspA - (G75S/G77S/L88K/D90N/D91N/I105E/D118R/Q126R/D134R/E139K)8	X	1	13
12	MspA - (G75S/G77S/L88N/D90N/D91N/D118G/Q126R/D134R/E139K)8	X	1	14

TABLE 3-continued

Entry	Mutant MspA Nanopore (SEQ ID NO: 2 with the specified mutations)	DNA Construct	Final Mg^{2+} concentration (mM)	Figure Number
13	MspA - (G75S/G77S/L88N/D90N/ D91N/D118N/Q126R/D134R/E139K)8	X	1	15
14	MspA - (G75S/G77S/L88N/D90N/ D91N/D118R/D134R/E139K)8	X	1	16
19	MspA - (G75S/G77S/L88K/D90N/D9IN/ N108E/D118R/Q126R/D134R/E139K)8	X	1	17
21	MspA - (G75S/G77S/L88N/D90N/D9IN/ T95E/P98K/D118R/Q126R/D134R/E139K)8	Y	2	18
22	MspA - (G75S/G77S/L88N/D90N/ D91N/D93N/D118R/Q126R/D134R/E139K)8	X	1	19

[0349] Entry 22 (FIG. **19**) shows the MspA mutant MspA—(G75S/G77 S/L88N/D90N/D91N/D93N/D118R/Q126R/D134R/E139K)8 which has substituted the aspartic acid at position 93 with a glycine. Helicase-controlled DNA movement was observed when DNA translocated through this mutant. However, entry 1 (FIG. **3**) has position 93 changed back to an aspartic acid MspA—(G75S/G77 S/L88N/D90N/D91N/D118R/Q126R/D134R/E139K)8 (which is the same amino acid that is present in the wild-type MspA) and this also exhibited helicase-controlled DNA movement. Therefore, it was possible to put the aspartic acid back into the nanopore amino acid sequence and helicase-controlled DNA movement was still observed.

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 29

<210> SEQ ID NO 1
<211> LENGTH: 552
<212> TYPE: DNA
<213> ORGANISM: Mycobacterium smegmatis

<400> SEQUENCE: 1

ggcctggata acgaacttag cctggtggac ggccaagatc gcacgctgac ggtgcaacaa 60

tgggatacct tcctgaatgg tgtgtttccg ctggatcgta accgcctgac ccgtgaatgg 120

tttcattccg gtcgcgcaaa atatatcgtc gcaggcccgg gtgctgacga attcgaaggc 180

acgctggaac tgggttatca gattggcttt ccgtggtcac tgggcgttgg tatcaacttc 240

tcgtacacca cgccgaatat tctgatcgat gacggtgata ttaccgcacc gccgtttggc 300

ctgaacagcg tgattacgcc gaacctgttt ccgggtgtta gcatctctgc cgatctgggc 360

aacggtccgg gcattcaaga agtggcaacc tttagtgtgg acgtttccgg cgctgaaggc 420

ggtgtcgcgg tgtctaatgc ccacggtacc gttacgggcg cggccggcgg tgtcctgctg 480

cgtccgttcg cgcgcctgat tgcgagcacc ggcgactctg ttacgaccta tggcgaaccg 540

tggaatatga ac 552

<210> SEQ ID NO 2
<211> LENGTH: 184
<212> TYPE: PRT
<213> ORGANISM: Mycobacterium smegmatis

<400> SEQUENCE: 2

Gly Leu Asp Asn Glu Leu Ser Leu Val Asp Gly Gln Asp Arg Thr Leu
1 5 10 15

Thr Val Gln Gln Trp Asp Thr Phe Leu Asn Gly Val Phe Pro Leu Asp
20 25 30

Arg Asn Arg Leu Thr Arg Glu Trp Phe His Ser Gly Arg Ala Lys Tyr
35 40 45

-continued

Ile Val Ala Gly Pro Gly Ala Asp Glu Phe Glu Gly Thr Leu Glu Leu
50 55 60

Gly Tyr Gln Ile Gly Phe Pro Trp Ser Leu Gly Val Gly Ile Asn Phe
65 70 75 80

Ser Tyr Thr Thr Pro Asn Ile Leu Ile Asp Asp Gly Asp Ile Thr Ala
85 90 95

Pro Pro Phe Gly Leu Asn Ser Val Ile Thr Pro Asn Leu Phe Pro Gly
100 105 110

Val Ser Ile Ser Ala Asp Leu Gly Asn Gly Pro Gly Ile Gln Glu Val
115 120 125

Ala Thr Phe Ser Val Asp Val Ser Gly Ala Glu Gly Gly Val Ala Val
130 135 140

Ser Asn Ala His Gly Thr Val Thr Gly Ala Ala Gly Gly Val Leu Leu
145 150 155 160

Arg Pro Phe Ala Arg Leu Ile Ala Ser Thr Gly Asp Ser Val Thr Thr
165 170 175

Tyr Gly Glu Pro Trp Asn Met Asn
180

<210> SEQ ID NO 3
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic Polynucleotide

<400> SEQUENCE: 3

ttttttttt ttttttttt 20

<210> SEQ ID NO 4
<211> LENGTH: 3595
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic Polynucleotide

<400> SEQUENCE: 4

gccatcagat tgtgtttgtt agtcgctgcc atcagattgt gtttgttagt cgctttttt 60

ttttggaatt ttttttttgg aatttttttt ttgcgctaac aacctcctgc cgttttgccc 120

gtgcatatcg gtcacgaaca aatctgatta ctaaacacag tagcctggat ttgttctatc 180

agtaatcgac cttattccta attaaataga gcaaatcccc ttattggggg taagacatga 240

agatgccaga aaaacatgac ctgttggccg ccattctcgc ggcaaaggaa caaggcatcg 300

gggcaatcct tgcgtttgca atggcgtacc ttcgcggcag atataatggc ggtgcgttta 360

caaaaacagt aatcgacgca acgatgtgcg ccattatcgc ctagttcatt cgtgaccttc 420

tcgacttcgc cggactaagt agcaatctcg cttatataac gagcgtgttt atcggctaca 480

tcggtactga ctcgattggt tcgcttatca aacgcttcgc tgctaaaaaa gccggagtag 540

aagatggtag aaatcaataa tcaacgtaag gcgttcctcg atatgctggc gtggtcggag 600

ggaactgata acggacgtca gaaaaccaga aatcatggtt atgacgtcat tgtaggcgga 660

gagctattta ctgattactc cgatcaccct cgcaaacttg tcacgctaaa cccaaaactc 720

aaatcaacag gcgccggacg ctaccagctt ctttcccgtt ggtgggatgc ctaccgcaag 780

cagcttggcc tgaaagactt ctctccgaaa agtcaggacg ctgtggcatt gcagcagatt 840

-continued

aaggagcgtg gcgctttacc tatgattgat cgtggtgata tccgtcaggc aatcgaccgt 900

tgcagcaata tctgggcttc actgccgggc gctggttatg gtcagttcga gcataaggct 960

gacagcctga ttgcaaaatt caaagaagcg ggcggaacgg tcagagagat tgatgtatga 1020

gcagagtcac cgcgattatc tccgctctgg ttatctgcat catcgtctgc ctgtcatggg 1080

ctgttaatca ttaccgtgat aacgccatta cctacaaagc ccagcgcgac aaaaatgcca 1140

gagaactgaa gctggcgaac gcggcaatta ctgacatgca gatgcgtcag cgtgatgttg 1200

ctgcgctcga tgcaaaatac acgaaggagt tagctgatgc taaagctgaa aatgatgctc 1260

tgcgtgatga tgttgccgct ggtcgtcgtc ggttgcacat caaagcagtc tgtcagtcag 1320

tgcgtgaagc caccaccgcc tccggcgtgg ataatgcagc ctccccccga ctggcagaca 1380

ccgctgaacg ggattatttc accctcagag agaggctgat cactatgcaa aaacaactgg 1440

aaggaaccca gaagtatatt aatgagcagt gcagatagag ttgcccatat cgatgggcaa 1500

ctcatgcaat tattgtgagc aatacacacg cgcttccagc ggagtataaa tgcctaaagt 1560

aataaaaccg agcaatccat ttacgaatgt ttgctgggtt tctgttttaa caacattttc 1620

tgcgccgcca caaattttgg ctgcatcgac agttttcttc tgcccaattc cagaaacgaa 1680

gaaatgatgg gtgatggttt cctttggtgc tactgctgcc ggtttgtttt gaacagtaaa 1740

cgtctgttga gcacatcctg taataagcag ggccagcgca gtagcgagta gcatttttt 1800

catggtgtta ttcccgatgc tttttgaagt tcgcagaatc gtatgtgtag aaaattaaac 1860

aaaccctaaa caatgagttg aaatttcata ttgttaatat ttattaatgt atgtcaggtg 1920

cgatgaatcg tcattgtatt cccggattaa ctatgtccac agccctgacg gggaacttct 1980

ctgcgggagt gtccgggaat aattaaaacg atgcacacag ggtttagcgc gtacacgtat 2040

tgcattatgc caacgccccg gtgctgacac ggaagaaacc ggacgttatg atttagcgtg 2100

gaaagatttg tgtagtgttc tgaatgctct cagtaaatag taatgaatta tcaaaggtat 2160

agtaatatct tttatgttca tggatatttg taacccatcg gaaaactcct gctttagcaa 2220

gattttccct gtattgctga aatgtgattt ctcttgattt caacctatca taggacgttt 2280

ctataagatg cgtgtttctt gagaatttaa catttacaac ctttttaagt ccttttatta 2340

acacggtgtt atcgttttct aacacgatgt gaatattatc tgtggctaga tagtaaatat 2400

aatgtgagac gttgtgacgt tttagttcag aataaaacaa ttcacagtct aaatcttttc 2460

gcacttgatc gaatatttct ttaaaaatgg caacctgagc cattggtaaa accttccatg 2520

tgatacgagg gcgcgtagtt tgcattatcg tttttatcgt ttcaatctgg tctgacctcc 2580

ttgtgttttg ttgatgattt atgtcaaata ttaggaatgt tttcacttaa tagtattggt 2640

tgcgtaacaa agtgcggtcc tgctggcatt ctggagggaa atacaaccga cagatgtatg 2700

taaggccaac gtgctcaaat cttcatacag aaagatttga agtaatattt taaccgctag 2760

atgaagagca agcgcatgga gcgacaaaat gaataaagaa caatctgctg atgatccctc 2820

cgtggatctg attcgtgtaa aaaatatgct taatagcacc atttctatga gttaccctga 2880

tgttgtaatt gcatgtatag aacataaggt gtctctggaa gcattcagag caattgaggc 2940

agcgttggtg aagcacgata ataatatgaa ggattattcc ctggtggttg actgatcacc 3000

ataactgcta atcattcaaa ctatttagtc tgtgacagag ccaacacgca gtctgtcact 3060

gtcaggaaag tggtaaaact gcaactcaat tactgcaatg ccctcgtaat taagtgaatt 3120

-continued

tacaatatcg tcctgttcgg agggaagaac gcgggatgtt cattcttcat cacttttaat 3180

tgatgtatat gctctctttt ctgacgttag tctccgacgg caggcttcaa tgacccaggc 3240

tgagaaattc ccggaccctt tttgctcaag agcgatgtta atttgttcaa tcatttggtt 3300

aggaaagcgg atgttgcggg ttgttgttct gcgggttctg ttcttcgttg acatgaggtt 3360

gccccgtatt cagtgtcgct gatttgtatt gtctgaagtt gtttttacgt taagttgatg 3420

cagatcaatt aatacgatac ctgcgtcata attgattatt tgacgtggtt tgatggcctc 3480

cacgcacgtt gtgatatgta gatgataatc attatcactt tacgggtcct ttccggtgaa 3540

aaaaaaggta ccaaaaaaaa catcgtcgtg agtagtgaac cgtaagcagc gacgg 3595

<210> SEQ ID NO 5
<211> LENGTH: 184
<212> TYPE: PRT
<213> ORGANISM: Mycobacterium smegmatis

<400> SEQUENCE: 5

Gly Leu Asp Asn Glu Leu Ser Leu Val Asp Gly Gln Asp Arg Thr Leu
1 5 10 15

Thr Val Gln Gln Trp Asp Thr Phe Leu Asn Gly Val Phe Pro Leu Asp
20 25 30

Arg Asn Arg Leu Thr Arg Glu Trp Phe His Ser Gly Arg Ala Lys Tyr
35 40 45

Ile Val Ala Gly Pro Gly Ala Asp Glu Phe Glu Gly Thr Leu Glu Leu
50 55 60

Gly Tyr Gln Ile Gly Phe Pro Trp Ser Leu Gly Val Gly Ile Asn Phe
65 70 75 80

Ser Tyr Thr Thr Pro Asn Ile Leu Ile Asp Asp Gly Asp Ile Thr Ala
85 90 95

Pro Pro Phe Gly Leu Asn Ser Val Ile Thr Pro Asn Leu Phe Pro Gly
100 105 110

Val Ser Ile Ser Ala Asp Leu Gly Asn Gly Pro Gly Ile Gln Glu Val
115 120 125

Ala Thr Phe Ser Val Asp Val Ser Gly Pro Ala Gly Gly Val Ala Val
130 135 140

Ser Asn Ala His Gly Thr Val Thr Gly Ala Ala Gly Gly Val Leu Leu
145 150 155 160

Arg Pro Phe Ala Arg Leu Ile Ala Ser Thr Gly Asp Ser Val Thr Thr
165 170 175

Tyr Gly Glu Pro Trp Asn Met Asn
180

<210> SEQ ID NO 6
<211> LENGTH: 184
<212> TYPE: PRT
<213> ORGANISM: Mycobacterium smegmatis

<400> SEQUENCE: 6

Gly Leu Asp Asn Glu Leu Ser Leu Val Asp Gly Gln Asp Arg Thr Leu
1 5 10 15

Thr Val Gln Gln Trp Asp Thr Phe Leu Asn Gly Val Phe Pro Leu Asp
20 25 30

Arg Asn Arg Leu Thr Arg Glu Trp Phe His Ser Gly Arg Ala Lys Tyr

-continued

```
        35                  40                  45

Ile Val Ala Gly Pro Gly Ala Asp Glu Phe Glu Gly Thr Leu Glu Leu
    50                  55                  60

Gly Tyr Gln Ile Gly Phe Pro Trp Ser Leu Gly Val Gly Ile Asn Phe
65                  70                  75                  80

Ser Tyr Thr Thr Pro Asn Ile Leu Ile Asp Asp Gly Asp Ile Thr Gly
                85                  90                  95

Pro Pro Phe Gly Leu Glu Ser Val Ile Thr Pro Asn Leu Phe Pro Gly
            100                 105                 110

Val Ser Ile Ser Ala Asp Leu Gly Asn Gly Pro Gly Ile Gln Glu Val
        115                 120                 125

Ala Thr Phe Ser Val Asp Val Ser Gly Pro Ala Gly Gly Val Ala Val
    130                 135                 140

Ser Asn Ala His Gly Thr Val Thr Gly Ala Ala Gly Gly Val Leu Leu
145                 150                 155                 160

Arg Pro Phe Ala Arg Leu Ile Ala Ser Thr Gly Asp Ser Val Thr Thr
                165                 170                 175

Tyr Gly Glu Pro Trp Asn Met Asn
            180


<210> SEQ ID NO 7
<211> LENGTH: 183
<212> TYPE: PRT
<213> ORGANISM: Mycobacterium smegmatis

<400> SEQUENCE: 7

Val Asp Asn Gln Leu Ser Val Val Asp Gly Gln Gly Arg Thr Leu Thr
1               5                   10                  15

Val Gln Gln Ala Glu Thr Phe Leu Asn Gly Val Phe Pro Leu Asp Arg
            20                  25                  30

Asn Arg Leu Thr Arg Glu Trp Phe His Ser Gly Arg Ala Thr Tyr His
        35                  40                  45

Val Ala Gly Pro Gly Ala Asp Glu Phe Glu Gly Thr Leu Glu Leu Gly
    50                  55                  60

Tyr Gln Val Gly Phe Pro Trp Ser Leu Gly Val Gly Ile Asn Phe Ser
65                  70                  75                  80

Tyr Thr Thr Pro Asn Ile Leu Ile Asp Gly Gly Asp Ile Thr Gln Pro
                85                  90                  95

Pro Phe Gly Leu Asp Thr Ile Ile Thr Pro Asn Leu Phe Pro Gly Val
            100                 105                 110

Ser Ile Ser Ala Asp Leu Gly Asn Gly Pro Gly Ile Gln Glu Val Ala
        115                 120                 125

Thr Phe Ser Val Asp Val Lys Gly Ala Lys Gly Ala Val Ala Val Ser
    130                 135                 140

Asn Ala His Gly Thr Val Thr Gly Ala Ala Gly Gly Val Leu Leu Arg
145                 150                 155                 160

Pro Phe Ala Arg Leu Ile Ala Ser Thr Gly Asp Ser Val Thr Thr Tyr
                165                 170                 175

Gly Glu Pro Trp Asn Met Asn
            180


<210> SEQ ID NO 8
<211> LENGTH: 1830
```

-continued

<212> TYPE: DNA
<213> ORGANISM: Bacteriophage phi-29

<400> SEQUENCE: 8

atgaaacaca tgccgcgtaa aatgtatagc tgcgcgtttg aaaccacgac caaagtggaa 60

gattgtcgcg tttgggccta tggctacatg aacatcgaag atcattctga atacaaaatc 120

ggtaacagtc tggatgaatt tatggcatgg gtgctgaaag ttcaggcgga tctgtacttc 180

cacaacctga aatttgatgg cgcattcatt atcaactggc tggaacgtaa tggctttaaa 240

tggagcgcgg atggtctgcc gaacacgtat aataccatta tctctcgtat gggccagtgg 300

tatatgattg atatctgcct gggctacaaa ggtaaacgca aaattcatac cgtgatctat 360

gatagcctga aaaaactgcc gtttccggtg aagaaaattg cgaaagattt caaactgacg 420

gttctgaaag gcgatattga ttatcacaaa gaacgtccgg ttggttacaa aatcaccccg 480

gaagaatacg catacatcaa aaacgatatc cagatcatcg cagaagcgct gctgattcag 540

tttaaacagg gcctggatcg catgaccgcg ggcagtgata gcctgaaagg tttcaaagat 600

atcatcacga ccaaaaaatt caaaaaagtg ttcccgacgc tgagcctggg tctggataaa 660

gaagttcgtt atgcctaccg cggcggtttt acctggctga acgatcgttt caaagaaaaa 720

gaaattggcg agggtatggt gtttgatgtt aatagtctgt atccggcaca gatgtacagc 780

cgcctgctgc cgtatggcga accgatcgtg ttcgagggta aatatgtttg ggatgaagat 840

tacccgctgc atattcagca catccgttgt gaatttgaac tgaaagaagg ctatattccg 900

accattcaga tcaaacgtag tcgcttctat aagggtaacg aatacctgaa aagctctggc 960

ggtgaaatcg cggatctgtg gctgagtaac gtggatctgg aactgatgaa agaacactac 1020

gatctgtaca acgttgaata catcagcggc ctgaaattta aagccacgac cggtctgttc 1080

aaagatttca tcgataaatg gacctacatc aaaacgacct ctgaaggcgc gattaaacag 1140

ctggccaaac tgatgctgaa cagcctgtat ggcaaattcg cctctaatcc ggatgtgacc 1200

ggtaaagttc cgtacctgaa agaaaatggc gcactgggtt ttcgcctggg cgaagaagaa 1260

acgaaagatc cggtgtatac cccgatgggt gttttcatta cggcctgggc acgttacacg 1320

accatcaccg cggcccaggc atgctatgat cgcattatct actgtgatac cgattctatt 1380

catctgacgg gcaccgaaat cccggatgtg attaaagata tcgttgatcc gaaaaaactg 1440

ggttattggg cccacgaaag tacgtttaaa cgtgcaaaat acctgcgcca gaaaacctac 1500

atccaggata tctacatgaa agaagtggat ggcaaactgg ttgaaggttc tccggatgat 1560

tacaccgata tcaaattcag tgtgaaatgc gccggcatga cggataaaat caaaaaagaa 1620

gtgaccttcg aaaacttcaa agttggtttc agccgcaaaa tgaaaccgaa accggtgcag 1680

gttccgggcg gtgtggttct ggtggatgat acgtttacca ttaaatctgg cggtagtgcg 1740

tggagccatc cgcagttcga aaaaggcggt ggctctggtg gcggttctgg cggtagtgcc 1800

tggagccacc cgcagtttga aaaataataa 1830

<210> SEQ ID NO 9
<211> LENGTH: 608
<212> TYPE: PRT
<213> ORGANISM: Bacteriophage phi-29

<400> SEQUENCE: 9

Met Lys His Met Pro Arg Lys Met Tyr Ser Cys Ala Phe Glu Thr Thr

-continued

```
1               5                   10                  15

Thr Lys Val Glu Asp Cys Arg Val Trp Ala Tyr Gly Tyr Met Asn Ile
            20                  25                  30

Glu Asp His Ser Glu Tyr Lys Ile Gly Asn Ser Leu Asp Glu Phe Met
        35                  40                  45

Ala Trp Val Leu Lys Val Gln Ala Asp Leu Tyr Phe His Asn Leu Lys
    50                  55                  60

Phe Asp Gly Ala Phe Ile Ile Asn Trp Leu Glu Arg Asn Gly Phe Lys
65                  70                  75                  80

Trp Ser Ala Asp Gly Leu Pro Asn Thr Tyr Asn Thr Ile Ile Ser Arg
                85                  90                  95

Met Gly Gln Trp Tyr Met Ile Asp Ile Cys Leu Gly Tyr Lys Gly Lys
            100                 105                 110

Arg Lys Ile His Thr Val Ile Tyr Asp Ser Leu Lys Lys Leu Pro Phe
        115                 120                 125

Pro Val Lys Lys Ile Ala Lys Asp Phe Lys Leu Thr Val Leu Lys Gly
    130                 135                 140

Asp Ile Asp Tyr His Lys Glu Arg Pro Val Gly Tyr Lys Ile Thr Pro
145                 150                 155                 160

Glu Glu Tyr Ala Tyr Ile Lys Asn Asp Ile Gln Ile Ile Ala Glu Ala
                165                 170                 175

Leu Leu Ile Gln Phe Lys Gln Gly Leu Asp Arg Met Thr Ala Gly Ser
            180                 185                 190

Asp Ser Leu Lys Gly Phe Lys Asp Ile Ile Thr Thr Lys Lys Phe Lys
        195                 200                 205

Lys Val Phe Pro Thr Leu Ser Leu Gly Leu Asp Lys Glu Val Arg Tyr
    210                 215                 220

Ala Tyr Arg Gly Gly Phe Thr Trp Leu Asn Asp Arg Phe Lys Glu Lys
225                 230                 235                 240

Glu Ile Gly Glu Gly Met Val Phe Asp Val Asn Ser Leu Tyr Pro Ala
                245                 250                 255

Gln Met Tyr Ser Arg Leu Leu Pro Tyr Gly Glu Pro Ile Val Phe Glu
            260                 265                 270

Gly Lys Tyr Val Trp Asp Glu Asp Tyr Pro Leu His Ile Gln His Ile
        275                 280                 285

Arg Cys Glu Phe Glu Leu Lys Glu Gly Tyr Ile Pro Thr Ile Gln Ile
    290                 295                 300

Lys Arg Ser Arg Phe Tyr Lys Gly Asn Glu Tyr Leu Lys Ser Ser Gly
305                 310                 315                 320

Gly Glu Ile Ala Asp Leu Trp Leu Ser Asn Val Asp Leu Glu Leu Met
                325                 330                 335

Lys Glu His Tyr Asp Leu Tyr Asn Val Glu Tyr Ile Ser Gly Leu Lys
            340                 345                 350

Phe Lys Ala Thr Thr Gly Leu Phe Lys Asp Phe Ile Asp Lys Trp Thr
        355                 360                 365

Tyr Ile Lys Thr Thr Ser Glu Gly Ala Ile Lys Gln Leu Ala Lys Leu
    370                 375                 380

Met Leu Asn Ser Leu Tyr Gly Lys Phe Ala Ser Asn Pro Asp Val Thr
385                 390                 395                 400

Gly Lys Val Pro Tyr Leu Lys Glu Asn Gly Ala Leu Gly Phe Arg Leu
                405                 410                 415
```

-continued

```
Gly Glu Glu Glu Thr Lys Asp Pro Val Tyr Thr Pro Met Gly Val Phe
            420                 425                 430

Ile Thr Ala Trp Ala Arg Tyr Thr Thr Ile Thr Ala Ala Gln Ala Cys
        435                 440                 445

Tyr Asp Arg Ile Ile Tyr Cys Asp Thr Asp Ser Ile His Leu Thr Gly
    450                 455                 460

Thr Glu Ile Pro Asp Val Ile Lys Asp Ile Val Asp Pro Lys Lys Leu
465                 470                 475                 480

Gly Tyr Trp Ala His Glu Ser Thr Phe Lys Arg Ala Lys Tyr Leu Arg
                485                 490                 495

Gln Lys Thr Tyr Ile Gln Asp Ile Tyr Met Lys Glu Val Asp Gly Lys
            500                 505                 510

Leu Val Glu Gly Ser Pro Asp Asp Tyr Thr Asp Ile Lys Phe Ser Val
        515                 520                 525

Lys Cys Ala Gly Met Thr Asp Lys Ile Lys Lys Glu Val Thr Phe Glu
    530                 535                 540

Asn Phe Lys Val Gly Phe Ser Arg Lys Met Lys Pro Lys Pro Val Gln
545                 550                 555                 560

Val Pro Gly Gly Val Val Leu Val Asp Asp Thr Phe Thr Ile Lys Ser
                565                 570                 575

Gly Gly Ser Ala Trp Ser His Pro Gln Phe Glu Lys Gly Gly Gly Ser
            580                 585                 590

Gly Gly Gly Ser Gly Gly Ser Ala Trp Ser His Pro Gln Phe Glu Lys
        595                 600                 605


<210> SEQ ID NO 10
<211> LENGTH: 1390
<212> TYPE: DNA
<213> ORGANISM: Escherichia coli

<400> SEQUENCE: 10

atgatgaacg atggcaaaca gcagagcacc ttcctgtttc atgattatga aaccttcggt     60

acccatccgg ccctggatcg tccggcgcag tttgcggcca ttcgcaccga tagcgaattc    120

aatgtgattg gcgaaccgga agtgttttat tgcaaaccgg ccgatgatta tctgccgcag    180

ccgggtgcgg tgctgattac cggtattacc ccgcaggaag cgcgcgcgaa aggtgaaaac    240

gaagcggcgt ttgccgcgcg cattcatagc ctgtttaccg tgccgaaaac ctgcattctg    300

ggctataaca atgtgcgctt cgatgatgaa gttacccgta atatctttta tcgtaacttt    360

tatgatccgt atgcgtggag ctggcagcat gataacagcc gttgggatct gctggatgtg    420

atgcgcgcgt gctatgcgct gcgcccggaa ggcattaatt ggccggaaaa cgatgatggc    480

ctgccgagct ttcgtctgga acatctgacc aaagccaacg gcattgaaca tagcaatgcc    540

catgatgcga tggccgatgt ttatgcgacc attgcgatgg cgaaactggt taaaacccgt    600

cagccgcgcc tgtttgatta tctgtttacc caccgtaaca aacacaaact gatggcgctg    660

attgatgttc cgcagatgaa accgctggtg catgtgagcg gcatgtttgg cgcctggcgc    720

ggcaacacca gctgggtggc cccgctggcc tggcacccgg aaaatcgtaa cgccgtgatt    780

atggttgatc tggccggtga tattagcccg ctgctggaac tggatagcga taccctgcgt    840

gaacgcctgt ataccgccaa aaccgatctg ggcgataatg ccgccgtgcc ggtgaaactg    900

gttcacatta acaaatgccc ggtgctggcc caggcgaaca ccctgcgccc ggaagatgcg    960
```

-continued

gatcgtctgg gtattaatcg ccagcattgt ctggataatc tgaaaatcct gcgtgaaaac 1020

ccgcaggtgc gtgaaaaagt ggtggcgatc ttcgcggaag cggaaccgtt caccccgagc 1080

gataacgtgg atgcgcagct gtataacggc ttctttagcg atgccgatcg cgcggcgatg 1140

aaaatcgttc tggaaaccga accgcgcaat ctgccggcgc tggatattac ctttgttgat 1200

aaacgtattg aaaaactgct gtttaattat cgtgcgcgca attttccggg taccctggat 1260

tatgccgaac agcagcgttg gctggaacat cgtcgtcagg ttttcacccc ggaatttctg 1320

cagggttatg cggatgaact gcagatgctg gttcagcagt atgccgatga taaagaaaaa 1380

gtggcgctgc 1390

<210> SEQ ID NO 11
<211> LENGTH: 485
<212> TYPE: PRT
<213> ORGANISM: Escherichia coli

<400> SEQUENCE: 11

Met Met Asn Asp Gly Lys Gln Gln Ser Thr Phe Leu Phe His Asp Tyr
1 5 10 15

Glu Thr Phe Gly Thr His Pro Ala Leu Asp Arg Pro Ala Gln Phe Ala
20 25 30

Ala Ile Arg Thr Asp Ser Glu Phe Asn Val Ile Gly Glu Pro Glu Val
35 40 45

Phe Tyr Cys Lys Pro Ala Asp Asp Tyr Leu Pro Gln Pro Gly Ala Val
50 55 60

Leu Ile Thr Gly Ile Thr Pro Gln Glu Ala Arg Ala Lys Gly Glu Asn
65 70 75 80

Glu Ala Ala Phe Ala Ala Arg Ile His Ser Leu Phe Thr Val Pro Lys
85 90 95

Thr Cys Ile Leu Gly Tyr Asn Asn Val Arg Phe Asp Asp Glu Val Thr
100 105 110

Arg Asn Ile Phe Tyr Arg Asn Phe Tyr Asp Pro Tyr Ala Trp Ser Trp
115 120 125

Gln His Asp Asn Ser Arg Trp Asp Leu Leu Asp Val Met Arg Ala Cys
130 135 140

Tyr Ala Leu Arg Pro Glu Gly Ile Asn Trp Pro Glu Asn Asp Asp Gly
145 150 155 160

Leu Pro Ser Phe Arg Leu Glu His Leu Thr Lys Ala Asn Gly Ile Glu
165 170 175

His Ser Asn Ala His Asp Ala Met Ala Asp Val Tyr Ala Thr Ile Ala
180 185 190

Met Ala Lys Leu Val Lys Thr Arg Gln Pro Arg Leu Phe Asp Tyr Leu
195 200 205

Phe Thr His Arg Asn Lys His Lys Leu Met Ala Leu Ile Asp Val Pro
210 215 220

Gln Met Lys Pro Leu Val His Val Ser Gly Met Phe Gly Ala Trp Arg
225 230 235 240

Gly Asn Thr Ser Trp Val Ala Pro Leu Ala Trp His Pro Glu Asn Arg
245 250 255

Asn Ala Val Ile Met Val Asp Leu Ala Gly Asp Ile Ser Pro Leu Leu
260 265 270

-continued

```
Glu Leu Asp Ser Asp Thr Leu Arg Glu Arg Leu Tyr Thr Ala Lys Thr
        275                 280                 285

Asp Leu Gly Asp Asn Ala Ala Val Pro Val Lys Leu Val His Ile Asn
    290                 295                 300

Lys Cys Pro Val Leu Ala Gln Ala Asn Thr Leu Arg Pro Glu Asp Ala
305                 310                 315                 320

Asp Arg Leu Gly Ile Asn Arg Gln His Cys Leu Asp Asn Leu Lys Ile
                325                 330                 335

Leu Arg Glu Asn Pro Gln Val Arg Glu Lys Val Val Ala Ile Phe Ala
            340                 345                 350

Glu Ala Glu Pro Phe Thr Pro Ser Asp Asn Val Asp Ala Gln Leu Tyr
        355                 360                 365

Asn Gly Phe Phe Ser Asp Ala Asp Arg Ala Ala Met Lys Ile Val Leu
    370                 375                 380

Glu Thr Glu Pro Arg Asn Leu Pro Ala Leu Asp Ile Thr Phe Val Asp
385                 390                 395                 400

Lys Arg Ile Glu Lys Leu Leu Phe Asn Tyr Arg Ala Arg Asn Phe Pro
                405                 410                 415

Gly Thr Leu Asp Tyr Ala Glu Gln Gln Arg Trp Leu Glu His Arg Arg
            420                 425                 430

Gln Val Phe Thr Pro Glu Phe Leu Gln Gly Tyr Ala Asp Glu Leu Gln
        435                 440                 445

Met Leu Val Gln Gln Tyr Ala Asp Asp Lys Glu Lys Val Ala Leu Leu
    450                 455                 460

Lys Ala Leu Trp Gln Tyr Ala Glu Glu Ile Val Ser Gly Ser Gly His
465                 470                 475                 480

His His His His His
                485


<210> SEQ ID NO 12
<211> LENGTH: 804
<212> TYPE: DNA
<213> ORGANISM: Escherichia coli

<400> SEQUENCE: 12

atgaaatttg tctcttttaa tatcaacggc ctgcgcgcca gacctcacca gcttgaagcc     60

atcgtcgaaa agcaccaacc ggatgtgatt ggcctgcagg agacaaaagt tcatgacgat    120

atgtttccgc tcgaagaggt ggcgaagctc ggctacaacg tgttttatca cgggcagaaa    180

ggccattatg gcgtggcgct gctgaccaaa gagacgccga ttgccgtgcg tcgcggcttt    240

cccggtgacg acgaagaggc gcagcggcgg attattatgg cggaaatccc ctcactgctg    300

ggtaatgtca ccgtgatcaa cggttacttc ccgcagggtg aaagccgcga ccatccgata    360

aaattcccgg caaaagcgca gttttatcag aatctgcaaa actacctgga aaccgaactc    420

aaacgtgata atccggtact gattatgggc gatatgaata tcagccctac agatctggat    480

atcggcattg gcgaagaaaa ccgtaagcgc tggctgcgta ccggtaaatg ctctttcctg    540

ccggaagagc gcgaatggat ggacaggctg atgagctggg ggttggtcga taccttccgc    600

catgcgaatc cgcaaacagc agatcgtttc tcatggtttg attaccgctc aaaaggtttt    660

gacgataacc gtggtctgcg catcgacctg ctgctcgcca gccaaccgct ggcagaatgt    720

tgcgtagaaa ccggcatcga ctatgaaatc cgcagcatgg aaaaaccgtc cgatcacgcc    780
```

-continued

cccgtctggg cgaccttccg ccgc 804

<210> SEQ ID NO 13
<211> LENGTH: 268
<212> TYPE: PRT
<213> ORGANISM: Escherichia coli

<400> SEQUENCE: 13

Met Lys Phe Val Ser Phe Asn Ile Asn Gly Leu Arg Ala Arg Pro His
1 5 10 15

Gln Leu Glu Ala Ile Val Glu Lys His Gln Pro Asp Val Ile Gly Leu
20 25 30

Gln Glu Thr Lys Val His Asp Asp Met Phe Pro Leu Glu Glu Val Ala
35 40 45

Lys Leu Gly Tyr Asn Val Phe Tyr His Gly Gln Lys Gly His Tyr Gly
50 55 60

Val Ala Leu Leu Thr Lys Glu Thr Pro Ile Ala Val Arg Arg Gly Phe
65 70 75 80

Pro Gly Asp Asp Glu Glu Ala Gln Arg Arg Ile Ile Met Ala Glu Ile
85 90 95

Pro Ser Leu Leu Gly Asn Val Thr Val Ile Asn Gly Tyr Phe Pro Gln
100 105 110

Gly Glu Ser Arg Asp His Pro Ile Lys Phe Pro Ala Lys Ala Gln Phe
115 120 125

Tyr Gln Asn Leu Gln Asn Tyr Leu Glu Thr Glu Leu Lys Arg Asp Asn
130 135 140

Pro Val Leu Ile Met Gly Asp Met Asn Ile Ser Pro Thr Asp Leu Asp
145 150 155 160

Ile Gly Ile Gly Glu Glu Asn Arg Lys Arg Trp Leu Arg Thr Gly Lys
165 170 175

Cys Ser Phe Leu Pro Glu Glu Arg Glu Trp Met Asp Arg Leu Met Ser
180 185 190

Trp Gly Leu Val Asp Thr Phe Arg His Ala Asn Pro Gln Thr Ala Asp
195 200 205

Arg Phe Ser Trp Phe Asp Tyr Arg Ser Lys Gly Phe Asp Asp Asn Arg
210 215 220

Gly Leu Arg Ile Asp Leu Leu Leu Ala Ser Gln Pro Leu Ala Glu Cys
225 230 235 240

Cys Val Glu Thr Gly Ile Asp Tyr Glu Ile Arg Ser Met Glu Lys Pro
245 250 255

Ser Asp His Ala Pro Val Trp Ala Thr Phe Arg Arg
260 265

<210> SEQ ID NO 14
<211> LENGTH: 1275
<212> TYPE: DNA
<213> ORGANISM: Thermus thermophilus

<400> SEQUENCE: 14

atgtttcgtc gtaaagaaga tctggatccg ccgctggcac tgctgccgct gaaaggcctg 60

cgcgaagccg ccgcactgct ggaagaagcg ctgcgtcaag gtaaacgcat tcgtgttcac 120

ggcgactatg atgcggatgg cctgaccggc accgcgatcc tggttcgtgg tctggccgcc 180

ctgggtgcgg atgttcatcc gtttatcccg caccgcctgg aagaaggcta tggtgtcctg 240

-continued

atggaacgcg tcccggaaca tctggaagcc tcggacctgt ttctgaccgt tgactgcggc 300

attaccaacc atgcggaact gcgcgaactg ctggaaaatg gcgtggaagt cattgttacc 360

gatcatcata cgccgggcaa aacgccgccg ccgggtctgg tcgtgcatcc ggcgctgacg 420

ccggatctga aagaaaaacc gaccggcgca ggcgtggcgt ttctgctgct gtgggcactg 480

catgaacgcc tgggcctgcc gccgccgctg gaatacgcgg acctggcagc cgttggcacc 540

attgccgacg ttgccccgct gtggggttgg aatcgtgcac tggtgaaaga aggtctggca 600

cgcatcccgg cttcatcttg ggtgggcctg cgtctgctgg ctgaagccgt gggctatacc 660

ggcaaagcgg tcgaagtcgc tttccgcatc gcgccgcgca tcaatgcggc ttcccgcctg 720

ggcgaagcgg aaaaagccct gcgcctgctg ctgacggatg atgcggcaga agctcaggcg 780

ctggtcggcg aactgcaccg tctgaacgcc cgtcgtcaga ccctggaaga agcgatgctg 840

cgcaaactgc tgccgcaggc cgacccggaa gcgaaagcca tcgttctgct ggacccggaa 900

ggccatccgg gtgttatggg tattgtggcc tctcgcatcc tggaagcgac cctgcgcccg 960

gtctttctgg tggcccaggg caaaggcacc gtgcgttcgc tggctccgat ttccgccgtc 1020

gaagcactgc gcagcgcgga agatctgctg ctgcgttatg gtggtcataa agaagcggcg 1080

ggtttcgcaa tggatgaagc gctgtttccg gcgttcaaag cacgcgttga agcgtatgcc 1140

gcacgtttcc cggatccggt tcgtgaagtg gcactgctgg atctgctgcc ggaaccgggc 1200

ctgctgccgc aggtgttccg tgaactggca ctgctggaac cgtatggtga aggtaacccg 1260

gaaccgctgt tcctg 1275

<210> SEQ ID NO 15
<211> LENGTH: 425
<212> TYPE: PRT
<213> ORGANISM: Thermus thermophilus

<400> SEQUENCE: 15

Met Phe Arg Arg Lys Glu Asp Leu Asp Pro Pro Leu Ala Leu Leu Pro
1 5 10 15

Leu Lys Gly Leu Arg Glu Ala Ala Ala Leu Leu Glu Glu Ala Leu Arg
20 25 30

Gln Gly Lys Arg Ile Arg Val His Gly Asp Tyr Asp Ala Asp Gly Leu
35 40 45

Thr Gly Thr Ala Ile Leu Val Arg Gly Leu Ala Ala Leu Gly Ala Asp
50 55 60

Val His Pro Phe Ile Pro His Arg Leu Glu Glu Gly Tyr Gly Val Leu
65 70 75 80

Met Glu Arg Val Pro Glu His Leu Glu Ala Ser Asp Leu Phe Leu Thr
85 90 95

Val Asp Cys Gly Ile Thr Asn His Ala Glu Leu Arg Glu Leu Leu Glu
100 105 110

Asn Gly Val Glu Val Ile Val Thr Asp His His Thr Pro Gly Lys Thr
115 120 125

Pro Pro Pro Gly Leu Val Val His Pro Ala Leu Thr Pro Asp Leu Lys
130 135 140

Glu Lys Pro Thr Gly Ala Gly Val Ala Phe Leu Leu Leu Trp Ala Leu
145 150 155 160

His Glu Arg Leu Gly Leu Pro Pro Pro Leu Glu Tyr Ala Asp Leu Ala

-continued

```
               165                 170                 175

Ala Val Gly Thr Ile Ala Asp Val Ala Pro Leu Trp Gly Trp Asn Arg
            180                 185                 190

Ala Leu Val Lys Glu Gly Leu Ala Arg Ile Pro Ala Ser Ser Trp Val
        195                 200                 205

Gly Leu Arg Leu Leu Ala Glu Ala Val Gly Tyr Thr Gly Lys Ala Val
    210                 215                 220

Glu Val Ala Phe Arg Ile Ala Pro Arg Ile Asn Ala Ala Ser Arg Leu
225                 230                 235                 240

Gly Glu Ala Glu Lys Ala Leu Arg Leu Leu Leu Thr Asp Asp Ala Ala
                245                 250                 255

Glu Ala Gln Ala Leu Val Gly Glu Leu His Arg Leu Asn Ala Arg Arg
            260                 265                 270

Gln Thr Leu Glu Glu Ala Met Leu Arg Lys Leu Leu Pro Gln Ala Asp
        275                 280                 285

Pro Glu Ala Lys Ala Ile Val Leu Leu Asp Pro Glu Gly His Pro Gly
    290                 295                 300

Val Met Gly Ile Val Ala Ser Arg Ile Leu Glu Ala Thr Leu Arg Pro
305                 310                 315                 320

Val Phe Leu Val Ala Gln Gly Lys Gly Thr Val Arg Ser Leu Ala Pro
                325                 330                 335

Ile Ser Ala Val Glu Ala Leu Arg Ser Ala Glu Asp Leu Leu Leu Arg
            340                 345                 350

Tyr Gly Gly His Lys Glu Ala Ala Gly Phe Ala Met Asp Glu Ala Leu
        355                 360                 365

Phe Pro Ala Phe Lys Ala Arg Val Glu Ala Tyr Ala Ala Arg Phe Pro
    370                 375                 380

Asp Pro Val Arg Glu Val Ala Leu Leu Asp Leu Leu Pro Glu Pro Gly
385                 390                 395                 400

Leu Leu Pro Gln Val Phe Arg Glu Leu Ala Leu Leu Glu Pro Tyr Gly
                405                 410                 415

Glu Gly Asn Pro Glu Pro Leu Phe Leu
            420                 425


<210> SEQ ID NO 16
<211> LENGTH: 738
<212> TYPE: DNA
<213> ORGANISM: Bacteriophage lambda

<400> SEQUENCE: 16

tccggaagcg gctctggtag tggttctggc atgacaccgg acattatcct gcagcgtacc     60

gggatcgatg tgagagctgt cgaacagggg gatgatgcgt ggcacaaatt acggctcggc    120

gtcatcaccg cttcagaagt tcacaacgtg atagcaaaac cccgctccgg aaagaagtgg    180

cctgacatga aaatgtccta cttccacacc ctgcttgctg aggtttgcac cggtgtggct    240

ccggaagtta acgctaaagc actggcctgg ggaaaacagt acgagaacga cgccagaacc    300

ctgtttgaat tcacttccgg cgtgaatgtt actgaatccc cgatcatcta tcgcgacgaa    360

agtatgcgta ccgcctgctc tcccgatggt ttatgcagtg acggcaacgg ccttgaactg    420

aaatgcccgt ttacctcccg ggatttcatg aagttccggc tcggtggttt cgaggccata    480

aagtcagctt acatggccca ggtgcagtac agcatgtggg tgacgcgaaa aaatgcctgg    540
```

-continued

tactttgcca actatgaccc gcgtatgaag cgtgaaggcc tgcattatgt cgtgattgag 600

cgggatgaaa agtacatggc gagttttgac gagatcgtgc cggagttcat cgaaaaaatg 660

gacgaggcac tggctgaaat tggttttgta tttggggagc aatggcgatc tggctctggt 720

tccggcagcg gttccgga 738

<210> SEQ ID NO 17
<211> LENGTH: 226
<212> TYPE: PRT
<213> ORGANISM: Bacteriophage lambda

<400> SEQUENCE: 17

Met Thr Pro Asp Ile Ile Leu Gln Arg Thr Gly Ile Asp Val Arg Ala
1 5 10 15

Val Glu Gln Gly Asp Asp Ala Trp His Lys Leu Arg Leu Gly Val Ile
20 25 30

Thr Ala Ser Glu Val His Asn Val Ile Ala Lys Pro Arg Ser Gly Lys
35 40 45

Lys Trp Pro Asp Met Lys Met Ser Tyr Phe His Thr Leu Leu Ala Glu
50 55 60

Val Cys Thr Gly Val Ala Pro Glu Val Asn Ala Lys Ala Leu Ala Trp
65 70 75 80

Gly Lys Gln Tyr Glu Asn Asp Ala Arg Thr Leu Phe Glu Phe Thr Ser
85 90 95

Gly Val Asn Val Thr Glu Ser Pro Ile Ile Tyr Arg Asp Glu Ser Met
100 105 110

Arg Thr Ala Cys Ser Pro Asp Gly Leu Cys Ser Asp Gly Asn Gly Leu
115 120 125

Glu Leu Lys Cys Pro Phe Thr Ser Arg Asp Phe Met Lys Phe Arg Leu
130 135 140

Gly Gly Phe Glu Ala Ile Lys Ser Ala Tyr Met Ala Gln Val Gln Tyr
145 150 155 160

Ser Met Trp Val Thr Arg Lys Asn Ala Trp Tyr Phe Ala Asn Tyr Asp
165 170 175

Pro Arg Met Lys Arg Glu Gly Leu His Tyr Val Val Ile Glu Arg Asp
180 185 190

Glu Lys Tyr Met Ala Ser Phe Asp Glu Ile Val Pro Glu Phe Ile Glu
195 200 205

Lys Met Asp Glu Ala Leu Ala Glu Ile Gly Phe Val Phe Gly Glu Gln
210 215 220

Trp Arg
225

<210> SEQ ID NO 18
<211> LENGTH: 760
<212> TYPE: PRT
<213> ORGANISM: Methanococcoides burtonii

<400> SEQUENCE: 18

Met Met Ile Arg Glu Leu Asp Ile Pro Arg Asp Ile Ile Gly Phe Tyr
1 5 10 15

Glu Asp Ser Gly Ile Lys Glu Leu Tyr Pro Pro Gln Ala Glu Ala Ile
20 25 30

Glu Met Gly Leu Leu Glu Lys Lys Asn Leu Leu Ala Ala Ile Pro Thr

-continued

```
        35                  40                  45

Ala Ser Gly Lys Thr Leu Leu Ala Glu Leu Ala Met Ile Lys Ala Ile
    50                  55                  60

Arg Glu Gly Gly Lys Ala Leu Tyr Ile Val Pro Leu Arg Ala Leu Ala
65                  70                  75                  80

Ser Glu Lys Phe Glu Arg Phe Lys Glu Leu Ala Pro Phe Gly Ile Lys
                85                  90                  95

Val Gly Ile Ser Thr Gly Asp Leu Asp Ser Arg Ala Asp Trp Leu Gly
            100                 105                 110

Val Asn Asp Ile Ile Val Ala Thr Ser Glu Lys Thr Asp Ser Leu Leu
        115                 120                 125

Arg Asn Gly Thr Ser Trp Met Asp Glu Ile Thr Thr Val Val Val Asp
    130                 135                 140

Glu Ile His Leu Leu Asp Ser Lys Asn Arg Gly Pro Thr Leu Glu Val
145                 150                 155                 160

Thr Ile Thr Lys Leu Met Arg Leu Asn Pro Asp Val Gln Val Val Ala
                165                 170                 175

Leu Ser Ala Thr Val Gly Asn Ala Arg Glu Met Ala Asp Trp Leu Gly
            180                 185                 190

Ala Ala Leu Val Leu Ser Glu Trp Arg Pro Thr Asp Leu His Glu Gly
        195                 200                 205

Val Leu Phe Gly Asp Ala Ile Asn Phe Pro Gly Ser Gln Lys Lys Ile
    210                 215                 220

Asp Arg Leu Glu Lys Asp Asp Ala Val Asn Leu Val Leu Asp Thr Ile
225                 230                 235                 240

Lys Ala Glu Gly Gln Cys Leu Val Phe Glu Ser Ser Arg Arg Asn Cys
                245                 250                 255

Ala Gly Phe Ala Lys Thr Ala Ser Ser Lys Val Ala Lys Ile Leu Asp
            260                 265                 270

Asn Asp Ile Met Ile Lys Leu Ala Gly Ile Ala Glu Glu Val Glu Ser
        275                 280                 285

Thr Gly Glu Thr Asp Thr Ala Ile Val Leu Ala Asn Cys Ile Arg Lys
    290                 295                 300

Gly Val Ala Phe His His Ala Gly Leu Asn Ser Asn His Arg Lys Leu
305                 310                 315                 320

Val Glu Asn Gly Phe Arg Gln Asn Leu Ile Lys Val Ile Ser Ser Thr
                325                 330                 335

Pro Thr Leu Ala Ala Gly Leu Asn Leu Pro Ala Arg Arg Val Ile Ile
            340                 345                 350

Arg Ser Tyr Arg Arg Phe Asp Ser Asn Phe Gly Met Gln Pro Ile Pro
        355                 360                 365

Val Leu Glu Tyr Lys Gln Met Ala Gly Arg Ala Gly Arg Pro His Leu
    370                 375                 380

Asp Pro Tyr Gly Glu Ser Val Leu Leu Ala Lys Thr Tyr Asp Glu Phe
385                 390                 395                 400

Ala Gln Leu Met Glu Asn Tyr Val Glu Ala Asp Ala Glu Asp Ile Trp
                405                 410                 415

Ser Lys Leu Gly Thr Glu Asn Ala Leu Arg Thr His Val Leu Ser Thr
            420                 425                 430

Ile Val Asn Gly Phe Ala Ser Thr Arg Gln Glu Leu Phe Asp Phe Phe
        435                 440                 445
```

-continued

```
Gly Ala Thr Phe Phe Ala Tyr Gln Gln Asp Lys Trp Met Leu Glu Glu
    450                 455                 460

Val Ile Asn Asp Cys Leu Glu Phe Leu Ile Asp Lys Ala Met Val Ser
465                 470                 475                 480

Glu Thr Glu Asp Ile Glu Asp Ala Ser Lys Leu Phe Leu Arg Gly Thr
                485                 490                 495

Arg Leu Gly Ser Leu Val Ser Met Leu Tyr Ile Asp Pro Leu Ser Gly
            500                 505                 510

Ser Lys Ile Val Asp Gly Phe Lys Asp Ile Gly Lys Ser Thr Gly Gly
        515                 520                 525

Asn Met Gly Ser Leu Glu Asp Asp Lys Gly Asp Asp Ile Thr Val Thr
    530                 535                 540

Asp Met Thr Leu Leu His Leu Val Cys Ser Thr Pro Asp Met Arg Gln
545                 550                 555                 560

Leu Tyr Leu Arg Asn Thr Asp Tyr Thr Ile Val Asn Glu Tyr Ile Val
                565                 570                 575

Ala His Ser Asp Glu Phe His Glu Ile Pro Asp Lys Leu Lys Glu Thr
            580                 585                 590

Asp Tyr Glu Trp Phe Met Gly Glu Val Lys Thr Ala Met Leu Leu Glu
        595                 600                 605

Glu Trp Val Thr Glu Val Ser Ala Glu Asp Ile Thr Arg His Phe Asn
    610                 615                 620

Val Gly Glu Gly Asp Ile His Ala Leu Ala Asp Thr Ser Glu Trp Leu
625                 630                 635                 640

Met His Ala Ala Ala Lys Leu Ala Glu Leu Leu Gly Val Glu Tyr Ser
                645                 650                 655

Ser His Ala Tyr Ser Leu Glu Lys Arg Ile Arg Tyr Gly Ser Gly Leu
            660                 665                 670

Asp Leu Met Glu Leu Val Gly Ile Arg Gly Val Gly Arg Val Arg Ala
        675                 680                 685

Arg Lys Leu Tyr Asn Ala Gly Phe Val Ser Val Ala Lys Leu Lys Gly
    690                 695                 700

Ala Asp Ile Ser Val Leu Ser Lys Leu Val Gly Pro Lys Val Ala Tyr
705                 710                 715                 720

Asn Ile Leu Ser Gly Ile Gly Val Arg Val Asn Asp Lys His Phe Asn
                725                 730                 735

Ser Ala Pro Ile Ser Ser Asn Thr Leu Asp Thr Leu Leu Asp Lys Asn
            740                 745                 750

Gln Lys Thr Phe Asn Asp Phe Gln
        755                 760


<210> SEQ ID NO 19
<211> LENGTH: 707
<212> TYPE: PRT
<213> ORGANISM: Cenarchaeum symbiosum

<400> SEQUENCE: 19

Met Arg Ile Ser Glu Leu Asp Ile Pro Arg Pro Ala Ile Glu Phe Leu
1               5                   10                  15

Glu Gly Glu Gly Tyr Lys Lys Leu Tyr Pro Pro Gln Ala Ala Ala Ala
            20                  25                  30

Lys Ala Gly Leu Thr Asp Gly Lys Ser Val Leu Val Ser Ala Pro Thr
```

-continued

```
        35                  40                  45

Ala Ser Gly Lys Thr Leu Ile Ala Ala Ile Ala Met Ile Ser His Leu
    50                  55                  60

Ser Arg Asn Arg Gly Lys Ala Val Tyr Leu Ser Pro Leu Arg Ala Leu
65                  70                  75                  80

Ala Ala Glu Lys Phe Ala Glu Phe Gly Lys Ile Gly Gly Ile Pro Leu
                85                  90                  95

Gly Arg Pro Val Arg Val Gly Val Ser Thr Gly Asp Phe Glu Lys Ala
            100                 105                 110

Gly Arg Ser Leu Gly Asn Asn Asp Ile Leu Val Leu Thr Asn Glu Arg
        115                 120                 125

Met Asp Ser Leu Ile Arg Arg Arg Pro Asp Trp Met Asp Glu Val Gly
    130                 135                 140

Leu Val Ile Ala Asp Glu Ile His Leu Ile Gly Asp Arg Ser Arg Gly
145                 150                 155                 160

Pro Thr Leu Glu Met Val Leu Thr Lys Leu Arg Gly Leu Arg Ser Ser
                165                 170                 175

Pro Gln Val Val Ala Leu Ser Ala Thr Ile Ser Asn Ala Asp Glu Ile
            180                 185                 190

Ala Gly Trp Leu Asp Cys Thr Leu Val His Ser Thr Trp Arg Pro Val
        195                 200                 205

Pro Leu Ser Glu Gly Val Tyr Gln Asp Gly Glu Val Ala Met Gly Asp
    210                 215                 220

Gly Ser Arg His Glu Val Ala Ala Thr Gly Gly Gly Pro Ala Val Asp
225                 230                 235                 240

Leu Ala Ala Glu Ser Val Ala Glu Gly Gly Gln Ser Leu Ile Phe Ala
                245                 250                 255

Asp Thr Arg Ala Arg Ser Ala Ser Leu Ala Ala Lys Ala Ser Ala Val
            260                 265                 270

Ile Pro Glu Ala Lys Gly Ala Asp Ala Ala Lys Leu Ala Ala Ala Ala
        275                 280                 285

Lys Lys Ile Ile Ser Ser Gly Gly Glu Thr Lys Leu Ala Lys Thr Leu
    290                 295                 300

Ala Glu Leu Val Glu Lys Gly Ala Ala Phe His His Ala Gly Leu Asn
305                 310                 315                 320

Gln Asp Cys Arg Ser Val Val Glu Glu Glu Phe Arg Ser Gly Arg Ile
                325                 330                 335

Arg Leu Leu Ala Ser Thr Pro Thr Leu Ala Ala Gly Val Asn Leu Pro
            340                 345                 350

Ala Arg Arg Val Val Ile Ser Ser Val Met Arg Tyr Asn Ser Ser Ser
        355                 360                 365

Gly Met Ser Glu Pro Ile Ser Ile Leu Glu Tyr Lys Gln Leu Cys Gly
    370                 375                 380

Arg Ala Gly Arg Pro Gln Tyr Asp Lys Ser Gly Glu Ala Ile Val Val
385                 390                 395                 400

Gly Gly Val Asn Ala Asp Glu Ile Phe Asp Arg Tyr Ile Gly Gly Glu
                405                 410                 415

Pro Glu Pro Ile Arg Ser Ala Met Val Asp Asp Arg Ala Leu Arg Ile
            420                 425                 430

His Val Leu Ser Leu Val Thr Thr Ser Pro Gly Ile Lys Glu Asp Asp
        435                 440                 445
```

-continued

```
Val Thr Glu Phe Phe Leu Gly Thr Leu Gly Gly Gln Gln Ser Gly Glu
    450                 455                 460

Ser Thr Val Lys Phe Ser Val Ala Val Ala Leu Arg Phe Leu Gln Glu
465                 470                 475                 480

Glu Gly Met Leu Gly Arg Arg Gly Gly Arg Leu Ala Ala Thr Lys Met
                485                 490                 495

Gly Arg Leu Val Ser Arg Leu Tyr Met Asp Pro Met Thr Ala Val Thr
            500                 505                 510

Leu Arg Asp Ala Val Gly Glu Ala Ser Pro Gly Arg Met His Thr Leu
        515                 520                 525

Gly Phe Leu His Leu Val Ser Glu Cys Ser Glu Phe Met Pro Arg Phe
    530                 535                 540

Ala Leu Arg Gln Lys Asp His Glu Val Ala Glu Met Met Leu Glu Ala
545                 550                 555                 560

Gly Arg Gly Glu Leu Leu Arg Pro Val Tyr Ser Tyr Glu Cys Gly Arg
                565                 570                 575

Gly Leu Leu Ala Leu His Arg Trp Ile Gly Glu Ser Pro Glu Ala Lys
            580                 585                 590

Leu Ala Glu Asp Leu Lys Phe Glu Ser Gly Asp Val His Arg Met Val
        595                 600                 605

Glu Ser Ser Gly Trp Leu Leu Arg Cys Ile Trp Glu Ile Ser Lys His
    610                 615                 620

Gln Glu Arg Pro Asp Leu Leu Gly Glu Leu Asp Val Leu Arg Ser Arg
625                 630                 635                 640

Val Ala Tyr Gly Ile Lys Ala Glu Leu Val Pro Leu Val Ser Ile Lys
                645                 650                 655

Gly Ile Gly Arg Val Arg Ser Arg Arg Leu Phe Arg Gly Gly Ile Lys
            660                 665                 670

Gly Pro Gly Asp Leu Ala Ala Val Pro Val Glu Arg Leu Ser Arg Val
        675                 680                 685

Glu Gly Ile Gly Ala Thr Leu Ala Asn Asn Ile Lys Ser Gln Leu Arg
    690                 695                 700

Lys Gly Gly
705


<210> SEQ ID NO 20
<211> LENGTH: 720
<212> TYPE: PRT
<213> ORGANISM: Thermococcus gammatolerans

<400> SEQUENCE: 20

Met Lys Val Asp Glu Leu Pro Val Asp Glu Arg Leu Lys Ala Val Leu
1               5                   10                  15

Lys Glu Arg Gly Ile Glu Glu Leu Tyr Pro Pro Gln Ala Glu Ala Leu
            20                  25                  30

Lys Ser Gly Ala Leu Glu Gly Arg Asn Leu Val Leu Ala Ile Pro Thr
        35                  40                  45

Ala Ser Gly Lys Thr Leu Val Ser Glu Ile Val Met Val Asn Lys Leu
    50                  55                  60

Ile Gln Glu Gly Gly Lys Ala Val Tyr Leu Val Pro Leu Lys Ala Leu
65                  70                  75                  80

Ala Glu Glu Lys Tyr Arg Glu Phe Lys Glu Trp Glu Lys Leu Gly Leu
```

-continued

```
                85                  90                  95

Lys Val Ala Ala Thr Thr Gly Asp Tyr Asp Ser Thr Asp Asp Trp Leu
            100                 105                 110

Gly Arg Tyr Asp Ile Ile Val Ala Thr Ala Glu Lys Phe Asp Ser Leu
        115                 120                 125

Leu Arg His Gly Ala Arg Trp Ile Asn Asp Val Lys Leu Val Val Ala
    130                 135                 140

Asp Glu Val His Leu Ile Gly Ser Tyr Asp Arg Gly Ala Thr Leu Glu
145                 150                 155                 160

Met Ile Leu Thr His Met Leu Gly Arg Ala Gln Ile Leu Ala Leu Ser
                165                 170                 175

Ala Thr Val Gly Asn Ala Glu Glu Leu Ala Glu Trp Leu Asp Ala Ser
            180                 185                 190

Leu Val Val Ser Asp Trp Arg Pro Val Gln Leu Arg Arg Gly Val Phe
        195                 200                 205

His Leu Gly Thr Leu Ile Trp Glu Asp Gly Lys Val Glu Ser Tyr Pro
    210                 215                 220

Glu Asn Trp Tyr Ser Leu Val Val Asp Ala Val Lys Arg Gly Lys Gly
225                 230                 235                 240

Ala Leu Val Phe Val Asn Thr Arg Arg Ser Ala Glu Lys Glu Ala Leu
                245                 250                 255

Ala Leu Ser Lys Leu Val Ser Ser His Leu Thr Lys Pro Glu Lys Arg
            260                 265                 270

Ala Leu Glu Ser Leu Ala Ser Gln Leu Glu Asp Asn Pro Thr Ser Glu
        275                 280                 285

Lys Leu Lys Arg Ala Leu Arg Gly Gly Val Ala Phe His His Ala Gly
    290                 295                 300

Leu Ser Arg Val Glu Arg Thr Leu Ile Glu Asp Ala Phe Arg Glu Gly
305                 310                 315                 320

Leu Ile Lys Val Ile Thr Ala Thr Pro Thr Leu Ser Ala Gly Val Asn
                325                 330                 335

Leu Pro Ser Phe Arg Val Ile Ile Arg Asp Thr Lys Arg Tyr Ala Gly
            340                 345                 350

Phe Gly Trp Thr Asp Ile Pro Val Leu Glu Ile Gln Gln Met Met Gly
        355                 360                 365

Arg Ala Gly Arg Pro Arg Tyr Asp Lys Tyr Gly Glu Ala Ile Ile Val
    370                 375                 380

Ala Arg Thr Asp Glu Pro Gly Lys Leu Met Glu Arg Tyr Ile Arg Gly
385                 390                 395                 400

Lys Pro Glu Lys Leu Phe Ser Met Leu Ala Asn Glu Gln Ala Phe Arg
                405                 410                 415

Ser Gln Val Leu Ala Leu Ile Thr Asn Phe Gly Ile Arg Ser Phe Pro
            420                 425                 430

Glu Leu Val Arg Phe Leu Glu Arg Thr Phe Tyr Ala His Gln Arg Lys
        435                 440                 445

Asp Leu Ser Ser Leu Glu Tyr Lys Ala Lys Glu Val Val Tyr Phe Leu
    450                 455                 460

Ile Glu Asn Glu Phe Ile Asp Leu Asp Leu Glu Asp Arg Phe Ile Pro
465                 470                 475                 480

Leu Pro Phe Gly Lys Arg Thr Ser Gln Leu Tyr Ile Asp Pro Leu Thr
                485                 490                 495
```

-continued

```
Ala Lys Lys Phe Lys Asp Ala Phe Pro Ala Ile Glu Arg Asn Pro Asn
            500                 505                 510

Pro Phe Gly Ile Phe Gln Leu Ile Ala Ser Thr Pro Asp Met Ala Thr
        515                 520                 525

Leu Thr Ala Arg Arg Arg Glu Met Glu Asp Tyr Leu Asp Leu Ala Tyr
    530                 535                 540

Glu Leu Glu Asp Lys Leu Tyr Ala Ser Ile Pro Tyr Tyr Glu Asp Ser
545                 550                 555                 560

Arg Phe Gln Gly Phe Leu Gly Gln Val Lys Thr Ala Lys Val Leu Leu
                565                 570                 575

Asp Trp Ile Asn Glu Val Pro Glu Ala Arg Ile Tyr Glu Thr Tyr Ser
            580                 585                 590

Ile Asp Pro Gly Asp Leu Tyr Arg Leu Leu Glu Leu Ala Asp Trp Leu
        595                 600                 605

Met Tyr Ser Leu Ile Glu Leu Tyr Lys Leu Phe Glu Pro Lys Glu Glu
    610                 615                 620

Ile Leu Asn Tyr Leu Arg Asp Leu His Leu Arg Leu Arg His Gly Val
625                 630                 635                 640

Arg Glu Glu Leu Leu Glu Leu Val Arg Leu Pro Asn Ile Gly Arg Lys
                645                 650                 655

Arg Ala Arg Ala Leu Tyr Asn Ala Gly Phe Arg Ser Val Glu Ala Ile
            660                 665                 670

Ala Asn Ala Lys Pro Ala Glu Leu Leu Ala Val Glu Gly Ile Gly Ala
        675                 680                 685

Lys Ile Leu Asp Gly Ile Tyr Arg His Leu Gly Ile Glu Lys Arg Val
    690                 695                 700

Thr Glu Glu Lys Pro Lys Arg Lys Gly Thr Leu Glu Asp Phe Leu Arg
705                 710                 715                 720


<210> SEQ ID NO 21
<211> LENGTH: 799
<212> TYPE: PRT
<213> ORGANISM: Methanospirillum hungatei

<400> SEQUENCE: 21

Met Glu Ile Ala Ser Leu Pro Leu Pro Asp Ser Phe Ile Arg Ala Cys
1               5                   10                  15

His Ala Lys Gly Ile Arg Ser Leu Tyr Pro Pro Gln Ala Glu Cys Ile
            20                  25                  30

Glu Lys Gly Leu Leu Glu Gly Lys Asn Leu Leu Ile Ser Ile Pro Thr
        35                  40                  45

Ala Ser Gly Lys Thr Leu Leu Ala Glu Met Ala Met Trp Ser Arg Ile
    50                  55                  60

Ala Ala Gly Gly Lys Cys Leu Tyr Ile Val Pro Leu Arg Ala Leu Ala
65                  70                  75                  80

Ser Glu Lys Tyr Asp Glu Phe Ser Lys Lys Gly Val Ile Arg Val Gly
                85                  90                  95

Ile Ala Thr Gly Asp Leu Asp Arg Thr Asp Ala Tyr Leu Gly Glu Asn
            100                 105                 110

Asp Ile Ile Val Ala Thr Ser Glu Lys Thr Asp Ser Leu Leu Arg Asn
        115                 120                 125

Arg Thr Pro Trp Leu Ser Gln Ile Thr Cys Ile Val Leu Asp Glu Val
```

-continued

```
    130                 135                 140

His Leu Ile Gly Ser Glu Asn Arg Gly Ala Thr Leu Glu Met Val Ile
145                 150                 155                 160

Thr Lys Leu Arg Tyr Thr Asn Pro Val Met Gln Ile Ile Gly Leu Ser
                165                 170                 175

Ala Thr Ile Gly Asn Pro Ala Gln Leu Ala Glu Trp Leu Asp Ala Thr
            180                 185                 190

Leu Ile Thr Ser Thr Trp Arg Pro Val Asp Leu Arg Gln Gly Val Tyr
        195                 200                 205

Tyr Asn Gly Lys Ile Arg Phe Ser Asp Ser Glu Arg Pro Ile Gln Gly
    210                 215                 220

Lys Thr Lys His Asp Asp Leu Asn Leu Cys Leu Asp Thr Ile Glu Glu
225                 230                 235                 240

Gly Gly Gln Cys Leu Val Phe Val Ser Ser Arg Arg Asn Ala Glu Gly
                245                 250                 255

Phe Ala Lys Lys Ala Ala Gly Ala Leu Lys Ala Gly Ser Pro Asp Ser
            260                 265                 270

Lys Ala Leu Ala Gln Glu Leu Arg Arg Leu Arg Asp Arg Asp Glu Gly
        275                 280                 285

Asn Val Leu Ala Asp Cys Val Glu Arg Gly Ala Ala Phe His His Ala
    290                 295                 300

Gly Leu Ile Arg Gln Glu Arg Thr Ile Ile Glu Glu Gly Phe Arg Asn
305                 310                 315                 320

Gly Tyr Ile Glu Val Ile Ala Ala Thr Pro Thr Leu Ala Ala Gly Leu
                325                 330                 335

Asn Leu Pro Ala Arg Arg Val Ile Ile Arg Asp Tyr Asn Arg Phe Ala
            340                 345                 350

Ser Gly Leu Gly Met Val Pro Ile Pro Val Gly Glu Tyr His Gln Met
        355                 360                 365

Ala Gly Arg Ala Gly Arg Pro His Leu Asp Pro Tyr Gly Glu Ala Val
    370                 375                 380

Leu Leu Ala Lys Asp Ala Pro Ser Val Glu Arg Leu Phe Glu Thr Phe
385                 390                 395                 400

Ile Asp Ala Glu Ala Glu Arg Val Asp Ser Gln Cys Val Asp Asp Ala
                405                 410                 415

Ser Leu Cys Ala His Ile Leu Ser Leu Ile Ala Thr Gly Phe Ala His
            420                 425                 430

Asp Gln Glu Ala Leu Ser Ser Phe Met Glu Arg Thr Phe Tyr Phe Phe
        435                 440                 445

Gln His Pro Lys Thr Arg Ser Leu Pro Arg Leu Val Ala Asp Ala Ile
    450                 455                 460

Arg Phe Leu Thr Thr Ala Gly Met Val Glu Glu Arg Glu Asn Thr Leu
465                 470                 475                 480

Ser Ala Thr Arg Leu Gly Ser Leu Val Ser Arg Leu Tyr Leu Asn Pro
                485                 490                 495

Cys Thr Ala Arg Leu Ile Leu Asp Ser Leu Lys Ser Cys Lys Thr Pro
            500                 505                 510

Thr Leu Ile Gly Leu Leu His Val Ile Cys Val Ser Pro Asp Met Gln
        515                 520                 525

Arg Leu Tyr Leu Lys Ala Ala Asp Thr Gln Leu Leu Arg Thr Phe Leu
    530                 535                 540
```

-continued

```
Phe Lys His Lys Asp Asp Leu Ile Leu Pro Leu Pro Phe Glu Gln Glu
545                 550                 555                 560

Glu Glu Glu Leu Trp Leu Ser Gly Leu Lys Thr Ala Leu Val Leu Thr
                565                 570                 575

Asp Trp Ala Asp Glu Phe Ser Glu Gly Met Ile Glu Glu Arg Tyr Gly
            580                 585                 590

Ile Gly Ala Gly Asp Leu Tyr Asn Ile Val Asp Ser Gly Lys Trp Leu
        595                 600                 605

Leu His Gly Thr Glu Arg Leu Val Ser Val Glu Met Pro Glu Met Ser
    610                 615                 620

Gln Val Val Lys Thr Leu Ser Val Arg Val His His Gly Val Lys Ser
625                 630                 635                 640

Glu Leu Leu Pro Leu Val Ala Leu Arg Asn Ile Gly Arg Val Arg Ala
                645                 650                 655

Arg Thr Leu Tyr Asn Ala Gly Tyr Pro Asp Pro Glu Ala Val Ala Arg
            660                 665                 670

Ala Gly Leu Ser Thr Ile Ala Arg Ile Ile Gly Glu Gly Ile Ala Arg
        675                 680                 685

Gln Val Ile Asp Glu Ile Thr Gly Val Lys Arg Ser Gly Ile His Ser
    690                 695                 700

Ser Asp Asp Asp Tyr Gln Gln Lys Thr Pro Glu Leu Leu Thr Asp Ile
705                 710                 715                 720

Pro Gly Ile Gly Lys Lys Met Ala Glu Lys Leu Gln Asn Ala Gly Ile
                725                 730                 735

Ile Thr Val Ser Asp Leu Leu Thr Ala Asp Glu Val Leu Leu Ser Asp
            740                 745                 750

Val Leu Gly Ala Ala Arg Ala Arg Lys Val Leu Ala Phe Leu Ser Asn
        755                 760                 765

Ser Glu Lys Glu Asn Ser Ser Ser Asp Lys Thr Glu Glu Ile Pro Asp
    770                 775                 780

Thr Gln Lys Ile Arg Gly Gln Ser Ser Trp Glu Asp Phe Gly Cys
785                 790                 795


<210> SEQ ID NO 22
<211> LENGTH: 1756
<212> TYPE: PRT
<213> ORGANISM: Escherichia coli

<400> SEQUENCE: 22

Met Met Ser Ile Ala Gln Val Arg Ser Ala Gly Ser Ala Gly Asn Tyr
1               5                   10                  15

Tyr Thr Asp Lys Asp Asn Tyr Tyr Val Leu Gly Ser Met Gly Glu Arg
            20                  25                  30

Trp Ala Gly Lys Gly Ala Glu Gln Leu Gly Leu Gln Gly Ser Val Asp
        35                  40                  45

Lys Asp Val Phe Thr Arg Leu Leu Glu Gly Arg Leu Pro Asp Gly Ala
    50                  55                  60

Asp Leu Ser Arg Met Gln Asp Gly Ser Asn Lys His Arg Pro Gly Tyr
65                  70                  75                  80

Asp Leu Thr Phe Ser Ala Pro Lys Ser Val Ser Met Met Ala Met Leu
                85                  90                  95

Gly Gly Asp Lys Arg Leu Ile Asp Ala His Asn Gln Ala Val Asp Phe
```

-continued

```
           100                 105                 110

Ala Val Arg Gln Val Glu Ala Leu Ala Ser Thr Arg Val Met Thr Asp
        115                 120                 125

Gly Gln Ser Glu Thr Val Leu Thr Gly Asn Leu Val Met Ala Leu Phe
    130                 135                 140

Asn His Asp Thr Ser Arg Asp Gln Glu Pro Gln Leu His Thr His Ala
145                 150                 155                 160

Val Val Ala Asn Val Thr Gln His Asn Gly Glu Trp Lys Thr Leu Ser
                165                 170                 175

Ser Asp Lys Val Gly Lys Thr Gly Phe Ile Glu Asn Val Tyr Ala Asn
            180                 185                 190

Gln Ile Ala Phe Gly Arg Leu Tyr Arg Glu Lys Leu Lys Glu Gln Val
        195                 200                 205

Glu Ala Leu Gly Tyr Glu Thr Glu Val Val Gly Lys His Gly Met Trp
    210                 215                 220

Glu Met Pro Gly Val Pro Val Glu Ala Phe Ser Gly Arg Ser Gln Ala
225                 230                 235                 240

Ile Arg Glu Ala Val Gly Glu Asp Ala Ser Leu Lys Ser Arg Asp Val
                245                 250                 255

Ala Ala Leu Asp Thr Arg Lys Ser Lys Gln His Val Asp Pro Glu Ile
            260                 265                 270

Arg Met Ala Glu Trp Met Gln Thr Leu Lys Glu Thr Gly Phe Asp Ile
        275                 280                 285

Arg Ala Tyr Arg Asp Ala Ala Asp Gln Arg Thr Glu Ile Arg Thr Gln
    290                 295                 300

Ala Pro Gly Pro Ala Ser Gln Asp Gly Pro Asp Val Gln Gln Ala Val
305                 310                 315                 320

Thr Gln Ala Ile Ala Gly Leu Ser Glu Arg Lys Val Gln Phe Thr Tyr
                325                 330                 335

Thr Asp Val Leu Ala Arg Thr Val Gly Ile Leu Pro Pro Glu Asn Gly
            340                 345                 350

Val Ile Glu Arg Ala Arg Ala Gly Ile Asp Glu Ala Ile Ser Arg Glu
        355                 360                 365

Gln Leu Ile Pro Leu Asp Arg Glu Lys Gly Leu Phe Thr Ser Gly Ile
    370                 375                 380

His Val Leu Asp Glu Leu Ser Val Arg Ala Leu Ser Arg Asp Ile Met
385                 390                 395                 400

Lys Gln Asn Arg Val Thr Val His Pro Glu Lys Ser Val Pro Arg Thr
                405                 410                 415

Ala Gly Tyr Ser Asp Ala Val Ser Val Leu Ala Gln Asp Arg Pro Ser
            420                 425                 430

Leu Ala Ile Val Ser Gly Gln Gly Gly Ala Ala Gly Gln Arg Glu Arg
        435                 440                 445

Val Ala Glu Leu Val Met Met Ala Arg Glu Gln Gly Arg Glu Val Gln
    450                 455                 460

Ile Ile Ala Ala Asp Arg Arg Ser Gln Met Asn Leu Lys Gln Asp Glu
465                 470                 475                 480

Arg Leu Ser Gly Glu Leu Ile Thr Gly Arg Arg Gln Leu Leu Glu Gly
                485                 490                 495

Met Ala Phe Thr Pro Gly Ser Thr Val Ile Val Asp Gln Gly Glu Lys
            500                 505                 510
```

-continued

```
Leu Ser Leu Lys Glu Thr Leu Thr Leu Leu Asp Gly Ala Ala Arg His
        515                 520                 525

Asn Val Gln Val Leu Ile Thr Asp Ser Gly Gln Arg Thr Gly Thr Gly
    530                 535                 540

Ser Ala Leu Met Ala Met Lys Asp Ala Gly Val Asn Thr Tyr Arg Trp
545                 550                 555                 560

Gln Gly Gly Glu Gln Arg Pro Ala Thr Ile Ile Ser Glu Pro Asp Arg
                565                 570                 575

Asn Val Arg Tyr Ala Arg Leu Ala Gly Asp Phe Ala Ala Ser Val Lys
            580                 585                 590

Ala Gly Glu Glu Ser Val Ala Gln Val Ser Gly Val Arg Glu Gln Ala
        595                 600                 605

Ile Leu Thr Gln Ala Ile Arg Ser Glu Leu Lys Thr Gln Gly Val Leu
    610                 615                 620

Gly His Pro Glu Val Thr Met Thr Ala Leu Ser Pro Val Trp Leu Asp
625                 630                 635                 640

Ser Arg Ser Arg Tyr Leu Arg Asp Met Tyr Arg Pro Gly Met Val Met
                645                 650                 655

Glu Gln Trp Asn Pro Glu Thr Arg Ser His Asp Arg Tyr Val Ile Asp
            660                 665                 670

Arg Val Thr Ala Gln Ser His Ser Leu Thr Leu Arg Asp Ala Gln Gly
        675                 680                 685

Glu Thr Gln Val Val Arg Ile Ser Ser Leu Asp Ser Ser Trp Ser Leu
    690                 695                 700

Phe Arg Pro Glu Lys Met Pro Val Ala Asp Gly Glu Arg Leu Arg Val
705                 710                 715                 720

Thr Gly Lys Ile Pro Gly Leu Arg Val Ser Gly Gly Asp Arg Leu Gln
                725                 730                 735

Val Ala Ser Val Ser Glu Asp Ala Met Thr Val Val Val Pro Gly Arg
            740                 745                 750

Ala Glu Pro Ala Ser Leu Pro Val Ser Asp Ser Pro Phe Thr Ala Leu
        755                 760                 765

Lys Leu Glu Asn Gly Trp Val Glu Thr Pro Gly His Ser Val Ser Asp
    770                 775                 780

Ser Ala Thr Val Phe Ala Ser Val Thr Gln Met Ala Met Asp Asn Ala
785                 790                 795                 800

Thr Leu Asn Gly Leu Ala Arg Ser Gly Arg Asp Val Arg Leu Tyr Ser
                805                 810                 815

Ser Leu Asp Glu Thr Arg Thr Ala Glu Lys Leu Ala Arg His Pro Ser
            820                 825                 830

Phe Thr Val Val Ser Glu Gln Ile Lys Ala Arg Ala Gly Glu Thr Leu
        835                 840                 845

Leu Glu Thr Ala Ile Ser Leu Gln Lys Ala Gly Leu His Thr Pro Ala
    850                 855                 860

Gln Gln Ala Ile His Leu Ala Leu Pro Val Leu Glu Ser Lys Asn Leu
865                 870                 875                 880

Ala Phe Ser Met Val Asp Leu Leu Thr Glu Ala Lys Ser Phe Ala Ala
                885                 890                 895

Glu Gly Thr Gly Phe Thr Glu Leu Gly Gly Glu Ile Asn Ala Gln Ile
            900                 905                 910
```

-continued

```
Lys Arg Gly Asp Leu Leu Tyr Val Asp Val Ala Lys Gly Tyr Gly Thr
        915                 920                 925

Gly Leu Leu Val Ser Arg Ala Ser Tyr Glu Ala Glu Lys Ser Ile Leu
    930                 935                 940

Arg His Ile Leu Glu Gly Lys Glu Ala Val Thr Pro Leu Met Glu Arg
945                 950                 955                 960

Val Pro Gly Glu Leu Met Glu Thr Leu Thr Ser Gly Gln Arg Ala Ala
                965                 970                 975

Thr Arg Met Ile Leu Glu Thr Ser Asp Arg Phe Thr Val Val Gln Gly
            980                 985                 990

Tyr Ala Gly Val Gly Lys Thr Thr  Gln Phe Arg Ala Val  Met Ser Ala
        995                 1000                1005

Val Asn  Met Leu Pro Ala Ser  Glu Arg Pro Arg Val  Val Gly Leu
    1010                1015                1020

Gly Pro  Thr His Arg Ala Val  Gly Glu Met Arg Ser  Ala Gly Val
    1025                1030                1035

Asp Ala  Gln Thr Leu Ala Ser  Phe Leu His Asp Thr  Gln Leu Gln
    1040                1045                1050

Gln Arg  Ser Gly Glu Thr Pro  Asp Phe Ser Asn Thr  Leu Phe Leu
    1055                1060                1065

Leu Asp  Glu Ser Ser Met Val  Gly Asn Thr Glu Met  Ala Arg Ala
    1070                1075                1080

Tyr Ala  Leu Ile Ala Ala Gly  Gly Gly Arg Ala Val  Ala Ser Gly
    1085                1090                1095

Asp Thr  Asp Gln Leu Gln Ala  Ile Ala Pro Gly Gln  Ser Phe Arg
    1100                1105                1110

Leu Gln  Gln Thr Arg Ser Ala  Ala Asp Val Val Ile  Met Lys Glu
    1115                1120                1125

Ile Val  Arg Gln Thr Pro Glu  Leu Arg Glu Ala Val  Tyr Ser Leu
    1130                1135                1140

Ile Asn  Arg Asp Val Glu Arg  Ala Leu Ser Gly Leu  Glu Ser Val
    1145                1150                1155

Lys Pro  Ser Gln Val Pro Arg  Leu Glu Gly Ala Trp  Ala Pro Glu
    1160                1165                1170

His Ser  Val Thr Glu Phe Ser  His Ser Gln Glu Ala  Lys Leu Ala
    1175                1180                1185

Glu Ala  Gln Gln Lys Ala Met  Leu Lys Gly Glu Ala  Phe Pro Asp
    1190                1195                1200

Ile Pro  Met Thr Leu Tyr Glu  Ala Ile Val Arg Asp  Tyr Thr Gly
    1205                1210                1215

Arg Thr  Pro Glu Ala Arg Glu  Gln Thr Leu Ile Val  Thr His Leu
    1220                1225                1230

Asn Glu  Asp Arg Arg Val Leu  Asn Ser Met Ile His  Asp Ala Arg
    1235                1240                1245

Glu Lys  Ala Gly Glu Leu Gly  Lys Glu Gln Val Met  Val Pro Val
    1250                1255                1260

Leu Asn  Thr Ala Asn Ile Arg  Asp Gly Glu Leu Arg  Arg Leu Ser
    1265                1270                1275

Thr Trp  Glu Lys Asn Pro Asp  Ala Leu Ala Leu Val  Asp Asn Val
    1280                1285                1290

Tyr His  Arg Ile Ala Gly Ile  Ser Lys Asp Asp Gly  Leu Ile Thr
```

-continued

```
    1295                1300                1305

Leu Gln  Asp Ala Glu Gly Asn  Thr Arg Leu Ile Ser  Pro Arg Glu
    1310                1315                1320

Ala Val  Ala Glu Gly Val Thr  Leu Tyr Thr Pro Asp  Lys Ile Arg
    1325                1330                1335

Val Gly  Thr Gly Asp Arg Met  Arg Phe Thr Lys Ser  Asp Arg Glu
    1340                1345                1350

Arg Gly  Tyr Val Ala Asn Ser  Val Trp Thr Val Thr  Ala Val Ser
    1355                1360                1365

Gly Asp  Ser Val Thr Leu Ser  Asp Gly Gln Gln Thr  Arg Val Ile
    1370                1375                1380

Arg Pro  Gly Gln Glu Arg Ala  Glu Gln His Ile Asp  Leu Ala Tyr
    1385                1390                1395

Ala Ile  Thr Ala His Gly Ala  Gln Gly Ala Ser Glu  Thr Phe Ala
    1400                1405                1410

Ile Ala  Leu Glu Gly Thr Glu  Gly Asn Arg Lys Leu  Met Ala Gly
    1415                1420                1425

Phe Glu  Ser Ala Tyr Val Ala  Leu Ser Arg Met Lys  Gln His Val
    1430                1435                1440

Gln Val  Tyr Thr Asp Asn Arg  Gln Gly Trp Thr Asp  Ala Ile Asn
    1445                1450                1455

Asn Ala  Val Gln Lys Gly Thr  Ala His Asp Val Leu  Glu Pro Lys
    1460                1465                1470

Pro Asp  Arg Glu Val Met Asn  Ala Gln Arg Leu Phe  Ser Thr Ala
    1475                1480                1485

Arg Glu  Leu Arg Asp Val Ala  Ala Gly Arg Ala Val  Leu Arg Gln
    1490                1495                1500

Ala Gly  Leu Ala Gly Gly Asp  Ser Pro Ala Arg Phe  Ile Ala Pro
    1505                1510                1515

Gly Arg  Lys Tyr Pro Gln Pro  Tyr Val Ala Leu Pro  Ala Phe Asp
    1520                1525                1530

Arg Asn  Gly Lys Ser Ala Gly  Ile Trp Leu Asn Pro  Leu Thr Thr
    1535                1540                1545

Asp Asp  Gly Asn Gly Leu Arg  Gly Phe Ser Gly Glu  Gly Arg Val
    1550                1555                1560

Lys Gly  Ser Gly Asp Ala Gln  Phe Val Ala Leu Gln  Gly Ser Arg
    1565                1570                1575

Asn Gly  Glu Ser Leu Leu Ala  Asp Asn Met Gln Asp  Gly Val Arg
    1580                1585                1590

Ile Ala  Arg Asp Asn Pro Asp  Ser Gly Val Val Val  Arg Ile Ala
    1595                1600                1605

Gly Glu  Gly Arg Pro Trp Asn  Pro Gly Ala Ile Thr  Gly Gly Arg
    1610                1615                1620

Val Trp  Gly Asp Ile Pro Asp  Asn Ser Val Gln Pro  Gly Ala Gly
    1625                1630                1635

Asn Gly  Glu Pro Val Thr Ala  Glu Val Leu Ala Gln  Arg Gln Ala
    1640                1645                1650

Glu Glu  Ala Ile Arg Arg Glu  Thr Glu Arg Arg Ala  Asp Glu Ile
    1655                1660                1665

Val Arg  Lys Met Ala Glu Asn  Lys Pro Asp Leu Pro  Asp Gly Lys
    1670                1675                1680
```

-continued

```
Thr Glu  Leu Ala Val Arg Asp  Ile Ala Gly Gln Glu  Arg Asp Arg
    1685                 1690                 1695

Ser Ala  Ile Ser Glu Arg Glu  Thr Ala Leu Pro Glu  Ser Val Leu
    1700                 1705                 1710

Arg Glu  Ser Gln Arg Glu Arg  Glu Ala Val Arg Glu  Val Ala Arg
    1715                 1720                 1725

Glu Asn  Leu Leu Gln Glu Arg  Leu Gln Gln Met Glu  Arg Asp Met
    1730                 1735                 1740

Val Arg  Asp Leu Gln Lys Glu  Lys Thr Leu Gly Gly  Asp
    1745                 1750                 1755


<210> SEQ ID NO 23
<211> LENGTH: 726
<212> TYPE: PRT
<213> ORGANISM: Methanococcoides burtonii

<400> SEQUENCE: 23

Met Ser Asp Lys Pro Ala Phe Met Lys Tyr Phe Thr Gln Ser Ser Cys
1               5                  10                  15

Tyr Pro Asn Gln Gln Glu Ala Met Asp Arg Ile His Ser Ala Leu Met
            20                  25                  30

Gln Gln Gln Leu Val Leu Phe Glu Gly Ala Cys Gly Thr Gly Lys Thr
        35                  40                  45

Leu Ser Ala Leu Val Pro Ala Leu His Val Gly Lys Met Leu Gly Lys
    50                  55                  60

Thr Val Ile Ile Ala Thr Asn Val His Gln Gln Met Val Gln Phe Ile
65                  70                  75                  80

Asn Glu Ala Arg Asp Ile Lys Lys Val Gln Asp Val Lys Val Ala Val
                85                  90                  95

Ile Lys Gly Lys Thr Ala Met Cys Pro Gln Glu Ala Asp Tyr Glu Glu
            100                 105                 110

Cys Ser Val Lys Arg Glu Asn Thr Phe Glu Leu Met Glu Thr Glu Arg
        115                 120                 125

Glu Ile Tyr Leu Lys Arg Gln Glu Leu Asn Ser Ala Arg Asp Ser Tyr
    130                 135                 140

Lys Lys Ser His Asp Pro Ala Phe Val Thr Leu Arg Asp Glu Leu Ser
145                 150                 155                 160

Lys Glu Ile Asp Ala Val Glu Glu Lys Ala Arg Gly Leu Arg Asp Arg
                165                 170                 175

Ala Cys Asn Asp Leu Tyr Glu Val Leu Arg Ser Asp Ser Glu Lys Phe
            180                 185                 190

Arg Glu Trp Leu Tyr Lys Glu Val Arg Ser Pro Glu Glu Ile Asn Asp
        195                 200                 205

His Ala Ile Lys Asp Gly Met Cys Gly Tyr Glu Leu Val Lys Arg Glu
    210                 215                 220

Leu Lys His Ala Asp Leu Leu Ile Cys Asn Tyr His His Val Leu Asn
225                 230                 235                 240

Pro Asp Ile Phe Ser Thr Val Leu Gly Trp Ile Glu Lys Glu Pro Gln
                245                 250                 255

Glu Thr Ile Val Ile Phe Asp Glu Ala His Asn Leu Glu Ser Ala Ala
            260                 265                 270

Arg Ser His Ser Ser Leu Ser Leu Thr Glu His Ser Ile Glu Lys Ala
```

-continued

```
        275                 280                 285

Ile Thr Glu Leu Glu Ala Asn Leu Asp Leu Leu Ala Asp Asp Asn Ile
    290                 295                 300

His Asn Leu Phe Asn Ile Phe Leu Glu Val Ile Ser Asp Thr Tyr Asn
305                 310                 315                 320

Ser Arg Phe Lys Phe Gly Glu Arg Glu Arg Val Arg Lys Asn Trp Tyr
                325                 330                 335

Asp Ile Arg Ile Ser Asp Pro Tyr Glu Arg Asn Asp Ile Val Arg Gly
            340                 345                 350

Lys Phe Leu Arg Gln Ala Lys Gly Asp Phe Gly Glu Lys Asp Asp Ile
        355                 360                 365

Gln Ile Leu Leu Ser Glu Ala Ser Glu Leu Gly Ala Lys Leu Asp Glu
    370                 375                 380

Thr Tyr Arg Asp Gln Tyr Lys Lys Gly Leu Ser Ser Val Met Lys Arg
385                 390                 395                 400

Ser His Ile Arg Tyr Val Ala Asp Phe Met Ser Ala Tyr Ile Glu Leu
                405                 410                 415

Ser His Asn Leu Asn Tyr Tyr Pro Ile Leu Asn Val Arg Arg Asp Met
            420                 425                 430

Asn Asp Glu Ile Tyr Gly Arg Val Glu Leu Phe Thr Cys Ile Pro Lys
        435                 440                 445

Asn Val Thr Glu Pro Leu Phe Asn Ser Leu Phe Ser Val Ile Leu Met
    450                 455                 460

Ser Ala Thr Leu His Pro Phe Glu Met Val Lys Lys Thr Leu Gly Ile
465                 470                 475                 480

Thr Arg Asp Thr Cys Glu Met Ser Tyr Gly Thr Ser Phe Pro Glu Glu
                485                 490                 495

Lys Arg Leu Ser Ile Ala Val Ser Ile Pro Pro Leu Phe Ala Lys Asn
            500                 505                 510

Arg Asp Asp Arg His Val Thr Glu Leu Leu Glu Gln Val Leu Leu Asp
        515                 520                 525

Ser Ile Glu Asn Ser Lys Gly Asn Val Ile Leu Phe Phe Gln Ser Ala
    530                 535                 540

Phe Glu Ala Lys Arg Tyr Tyr Ser Lys Ile Glu Pro Leu Val Asn Val
545                 550                 555                 560

Pro Val Phe Leu Asp Glu Val Gly Ile Ser Ser Gln Asp Val Arg Glu
                565                 570                 575

Glu Phe Phe Ser Ile Gly Glu Glu Asn Gly Lys Ala Val Leu Leu Ser
            580                 585                 590

Tyr Leu Trp Gly Thr Leu Ser Glu Gly Ile Asp Tyr Arg Asp Gly Arg
        595                 600                 605

Gly Arg Thr Val Ile Ile Ile Gly Val Gly Tyr Pro Ala Leu Asn Asp
    610                 615                 620

Arg Met Asn Ala Val Glu Ser Ala Tyr Asp His Val Phe Gly Tyr Gly
625                 630                 635                 640

Ala Gly Trp Glu Phe Ala Ile Gln Val Pro Thr Ile Arg Lys Ile Arg
                645                 650                 655

Gln Ala Met Gly Arg Val Val Arg Ser Pro Thr Asp Tyr Gly Ala Arg
            660                 665                 670

Ile Leu Leu Asp Gly Arg Phe Leu Thr Asp Ser Lys Lys Arg Phe Gly
        675                 680                 685
```

-continued

```
Lys Phe Ser Val Phe Glu Val Phe Pro Pro Ala Glu Arg Ser Glu Phe
    690                 695                 700

Val Asp Val Asp Pro Glu Lys Val Lys Tyr Ser Leu Met Asn Phe Phe
705                 710                 715                 720

Met Asp Asn Asp Glu Gln
                725


<210> SEQ ID NO 24
<211> LENGTH: 439
<212> TYPE: PRT
<213> ORGANISM: Dickeya dadantii

<400> SEQUENCE: 24

Met Thr Phe Asp Asp Leu Thr Glu Gly Gln Lys Asn Ala Phe Asn Ile
1               5                   10                  15

Val Met Lys Ala Ile Lys Glu Lys Lys His His Val Thr Ile Asn Gly
            20                  25                  30

Pro Ala Gly Thr Gly Lys Thr Thr Leu Thr Lys Phe Ile Ile Glu Ala
        35                  40                  45

Leu Ile Ser Thr Gly Glu Thr Gly Ile Ile Leu Ala Ala Pro Thr His
    50                  55                  60

Ala Ala Lys Lys Ile Leu Ser Lys Leu Ser Gly Lys Glu Ala Ser Thr
65                  70                  75                  80

Ile His Ser Ile Leu Lys Ile Asn Pro Val Thr Tyr Glu Glu Asn Val
                85                  90                  95

Leu Phe Glu Gln Lys Glu Val Pro Asp Leu Ala Lys Cys Arg Val Leu
            100                 105                 110

Ile Cys Asp Glu Val Ser Met Tyr Asp Arg Lys Leu Phe Lys Ile Leu
        115                 120                 125

Leu Ser Thr Ile Pro Pro Trp Cys Thr Ile Ile Gly Ile Gly Asp Asn
    130                 135                 140

Lys Gln Ile Arg Pro Val Asp Pro Gly Glu Asn Thr Ala Tyr Ile Ser
145                 150                 155                 160

Pro Phe Phe Thr His Lys Asp Phe Tyr Gln Cys Glu Leu Thr Glu Val
                165                 170                 175

Lys Arg Ser Asn Ala Pro Ile Ile Asp Val Ala Thr Asp Val Arg Asn
            180                 185                 190

Gly Lys Trp Ile Tyr Asp Lys Val Val Asp Gly His Gly Val Arg Gly
        195                 200                 205

Phe Thr Gly Asp Thr Ala Leu Arg Asp Phe Met Val Asn Tyr Phe Ser
    210                 215                 220

Ile Val Lys Ser Leu Asp Asp Leu Phe Glu Asn Arg Val Met Ala Phe
225                 230                 235                 240

Thr Asn Lys Ser Val Asp Lys Leu Asn Ser Ile Ile Arg Lys Lys Ile
                245                 250                 255

Phe Glu Thr Asp Lys Asp Phe Ile Val Gly Glu Ile Ile Val Met Gln
            260                 265                 270

Glu Pro Leu Phe Lys Thr Tyr Lys Ile Asp Gly Lys Pro Val Ser Glu
        275                 280                 285

Ile Ile Phe Asn Asn Gly Gln Leu Val Arg Ile Ile Glu Ala Glu Tyr
    290                 295                 300

Thr Ser Thr Phe Val Lys Ala Arg Gly Val Pro Gly Glu Tyr Leu Ile
```

-continued

```
305                 310                 315                 320

Arg His Trp Asp Leu Thr Val Glu Thr Tyr Gly Asp Asp Glu Tyr Tyr
                325                 330                 335

Arg Glu Lys Ile Lys Ile Ile Ser Ser Asp Glu Glu Leu Tyr Lys Phe
            340                 345                 350

Asn Leu Phe Leu Gly Lys Thr Ala Glu Thr Tyr Lys Asn Trp Asn Lys
        355                 360                 365

Gly Gly Lys Ala Pro Trp Ser Asp Phe Trp Asp Ala Lys Ser Gln Phe
    370                 375                 380

Ser Lys Val Lys Ala Leu Pro Ala Ser Thr Phe His Lys Ala Gln Gly
385                 390                 395                 400

Met Ser Val Asp Arg Ala Phe Ile Tyr Thr Pro Cys Ile His Tyr Ala
                405                 410                 415

Asp Val Glu Leu Ala Gln Gln Leu Leu Tyr Val Gly Val Thr Arg Gly
            420                 425                 430

Arg Tyr Asp Val Phe Tyr Val
        435


<210> SEQ ID NO 25
<211> LENGTH: 970
<212> TYPE: PRT
<213> ORGANISM: Clostridium botulinum

<400> SEQUENCE: 25

Met Leu Ser Val Ala Asn Val Arg Ser Pro Ser Ala Ala Ala Ser Tyr
1               5                   10                  15

Phe Ala Ser Asp Asn Tyr Tyr Ala Ser Ala Asp Ala Asp Arg Ser Gly
            20                  25                  30

Gln Trp Ile Gly Asp Gly Ala Lys Arg Leu Gly Leu Glu Gly Lys Val
        35                  40                  45

Glu Ala Arg Ala Phe Asp Ala Leu Leu Arg Gly Glu Leu Pro Asp Gly
    50                  55                  60

Ser Ser Val Gly Asn Pro Gly Gln Ala His Arg Pro Gly Thr Asp Leu
65                  70                  75                  80

Thr Phe Ser Val Pro Lys Ser Trp Ser Leu Leu Ala Leu Val Gly Lys
                85                  90                  95

Asp Glu Arg Ile Ile Ala Ala Tyr Arg Glu Ala Val Val Glu Ala Leu
            100                 105                 110

His Trp Ala Glu Lys Asn Ala Ala Glu Thr Arg Val Val Glu Lys Gly
        115                 120                 125

Met Val Val Thr Gln Ala Thr Gly Asn Leu Ala Ile Gly Leu Phe Gln
    130                 135                 140

His Asp Thr Asn Arg Asn Gln Glu Pro Asn Leu His Phe His Ala Val
145                 150                 155                 160

Ile Ala Asn Val Thr Gln Gly Lys Asp Gly Lys Trp Arg Thr Leu Lys
                165                 170                 175

Asn Asp Arg Leu Trp Gln Leu Asn Thr Thr Leu Asn Ser Ile Ala Met
            180                 185                 190

Ala Arg Phe Arg Val Ala Val Glu Lys Leu Gly Tyr Glu Pro Gly Pro
        195                 200                 205

Val Leu Lys His Gly Asn Phe Glu Ala Arg Gly Ile Ser Arg Glu Gln
    210                 215                 220
```

-continued

```
Val Met Ala Phe Ser Thr Arg Arg Lys Glu Val Leu Glu Ala Arg Arg
225                 230                 235                 240

Gly Pro Gly Leu Asp Ala Gly Arg Ile Ala Ala Leu Asp Thr Arg Ala
                245                 250                 255

Ser Lys Glu Gly Ile Glu Asp Arg Ala Thr Leu Ser Lys Gln Trp Ser
            260                 265                 270

Glu Ala Ala Gln Ser Ile Gly Leu Asp Leu Lys Pro Leu Val Asp Arg
        275                 280                 285

Ala Arg Thr Lys Ala Leu Gly Gln Gly Met Glu Ala Thr Arg Ile Gly
    290                 295                 300

Ser Leu Val Glu Arg Gly Arg Ala Trp Leu Ser Arg Phe Ala Ala His
305                 310                 315                 320

Val Arg Gly Asp Pro Ala Asp Pro Leu Val Pro Pro Ser Val Leu Lys
                325                 330                 335

Gln Asp Arg Gln Thr Ile Ala Ala Ala Gln Ala Val Ala Ser Ala Val
            340                 345                 350

Arg His Leu Ser Gln Arg Glu Ala Ala Phe Glu Arg Thr Ala Leu Tyr
        355                 360                 365

Lys Ala Ala Leu Asp Phe Gly Leu Pro Thr Thr Ile Ala Asp Val Glu
    370                 375                 380

Lys Arg Thr Arg Ala Leu Val Arg Ser Gly Asp Leu Ile Ala Gly Lys
385                 390                 395                 400

Gly Glu His Lys Gly Trp Leu Ala Ser Arg Asp Ala Val Val Thr Glu
                405                 410                 415

Gln Arg Ile Leu Ser Glu Val Ala Ala Gly Lys Gly Asp Ser Ser Pro
            420                 425                 430

Ala Ile Thr Pro Gln Lys Ala Ala Ala Ser Val Gln Ala Ala Ala Leu
        435                 440                 445

Thr Gly Gln Gly Phe Arg Leu Asn Glu Gly Gln Leu Ala Ala Ala Arg
    450                 455                 460

Leu Ile Leu Ile Ser Lys Asp Arg Thr Ile Ala Val Gln Gly Ile Ala
465                 470                 475                 480

Gly Ala Gly Lys Ser Ser Val Leu Lys Pro Val Ala Glu Val Leu Arg
                485                 490                 495

Asp Glu Gly His Pro Val Ile Gly Leu Ala Ile Gln Asn Thr Leu Val
            500                 505                 510

Gln Met Leu Glu Arg Asp Thr Gly Ile Gly Ser Gln Thr Leu Ala Arg
        515                 520                 525

Phe Leu Gly Gly Trp Asn Lys Leu Leu Asp Asp Pro Gly Asn Val Ala
    530                 535                 540

Leu Arg Ala Glu Ala Gln Ala Ser Leu Lys Asp His Val Leu Val Leu
545                 550                 555                 560

Asp Glu Ala Ser Met Val Ser Asn Glu Asp Lys Glu Lys Leu Val Arg
                565                 570                 575

Leu Ala Asn Leu Ala Gly Val His Arg Leu Val Leu Ile Gly Asp Arg
            580                 585                 590

Lys Gln Leu Gly Ala Val Asp Ala Gly Lys Pro Phe Ala Leu Leu Gln
        595                 600                 605

Arg Ala Gly Ile Ala Arg Ala Glu Met Ala Thr Asn Leu Arg Ala Arg
    610                 615                 620

Asp Pro Val Val Arg Glu Ala Gln Ala Ala Ala Gln Ala Gly Asp Val
```

-continued

625 630 635 640

Arg Lys Ala Leu Arg His Leu Lys Ser His Thr Val Glu Ala Arg Gly
645 650 655

Asp Gly Ala Gln Val Ala Ala Glu Thr Trp Leu Ala Leu Asp Lys Glu
660 665 670

Thr Arg Ala Arg Thr Ser Ile Tyr Ala Ser Gly Arg Ala Ile Arg Ser
675 680 685

Ala Val Asn Ala Ala Val Gln Gln Gly Leu Leu Ala Ser Arg Glu Ile
690 695 700

Gly Pro Ala Lys Met Lys Leu Glu Val Leu Asp Arg Val Asn Thr Thr
705 710 715 720

Arg Glu Glu Leu Arg His Leu Pro Ala Tyr Arg Ala Gly Arg Val Leu
725 730 735

Glu Val Ser Arg Lys Gln Gln Ala Leu Gly Leu Phe Ile Gly Glu Tyr
740 745 750

Arg Val Ile Gly Gln Asp Arg Lys Gly Lys Leu Val Glu Val Glu Asp
755 760 765

Lys Arg Gly Lys Arg Phe Arg Phe Asp Pro Ala Arg Ile Arg Ala Gly
770 775 780

Lys Gly Asp Asp Asn Leu Thr Leu Leu Glu Pro Arg Lys Leu Glu Ile
785 790 795 800

His Glu Gly Asp Arg Ile Arg Trp Thr Arg Asn Asp His Arg Arg Gly
805 810 815

Leu Phe Asn Ala Asp Gln Ala Arg Val Val Glu Ile Ala Asn Gly Lys
820 825 830

Val Thr Phe Glu Thr Ser Lys Gly Asp Leu Val Glu Leu Lys Lys Asp
835 840 845

Asp Pro Met Leu Lys Arg Ile Asp Leu Ala Tyr Ala Leu Asn Val His
850 855 860

Met Ala Gln Gly Leu Thr Ser Asp Arg Gly Ile Ala Val Met Asp Ser
865 870 875 880

Arg Glu Arg Asn Leu Ser Asn Gln Lys Thr Phe Leu Val Thr Val Thr
885 890 895

Arg Leu Arg Asp His Leu Thr Leu Val Val Asp Ser Ala Asp Lys Leu
900 905 910

Gly Ala Ala Val Ala Arg Asn Lys Gly Glu Lys Ala Ser Ala Ile Glu
915 920 925

Val Thr Gly Ser Val Lys Pro Thr Ala Thr Lys Gly Ser Gly Val Asp
930 935 940

Gln Pro Lys Ser Val Glu Ala Asn Lys Ala Glu Lys Glu Leu Thr Arg
945 950 955 960

Ser Lys Ser Lys Thr Leu Asp Phe Gly Ile
965 970

<210> SEQ ID NO 26
<211> LENGTH: 46
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic Polynucleotide

<400> SEQUENCE: 26

ttttttttttt ttttttttttt ttttttttttt ttttttttttt tttttt 46

-continued

<210> SEQ ID NO 27
<211> LENGTH: 27
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic Polynucleotide

<400> SEQUENCE: 27

ggttgtttct gttggtgctg atattgc 27

<210> SEQ ID NO 28
<211> LENGTH: 3588
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic Polynucleotide

<400> SEQUENCE: 28

gccatcagat tgtgtttgtt agtcgctgcc atcagattgt gtttgttagt cgctttttt 60

ttttggaatt ttttttttgg aatttttttt ttgcgctaac aacctcctgc cgttttgccc 120

gtgcatatcg gtcacgaaca aatctgatta ctaaacacag tagcctggat ttgttctatc 180

agtaatcgac cttattccta attaaataga gcaaatcccc ttattggggg taagacatga 240

agatgccaga aaaacatgac ctgttggccg ccattctcgc ggcaaaggaa caaggcatcg 300

gggcaatcct tgcgtttgca atggcgtacc ttcgcggcag atataatggc ggtgcgttta 360

caaaaacagt aatcgacgca acgatgtgcg ccattatcgc ctagttcatt cgtgaccttc 420

tcgacttcgc cggactaagt agcaatctcg cttatataac gagcgtgttt atcggctaca 480

tcggtactga ctcgattggt tcgcttatca aacgcttcgc tgctaaaaaa gccggagtag 540

aagatggtag aaatcaataa tcaacgtaag gcgttcctcg atatgctggc gtggtcggag 600

ggaactgata acggacgtca gaaaaccaga aatcatggtt atgacgtcat tgtaggcgga 660

gagctattta ctgattactc cgatcaccct cgcaaacttg tcacgctaaa cccaaaactc 720

aaatcaacag gcgccggacg ctaccagctt ctttcccgtt ggtgggatgc ctaccgcaag 780

cagcttggcc tgaaagactt ctctccgaaa agtcaggacg ctgtggcatt gcagcagatt 840

aaggagcgtg gcgctttacc tatgattgat cgtggtgata tccgtcaggc aatcgaccgt 900

tgcagcaata tctgggcttc actgccgggc gctggttatg gtcagttcga gcataaggct 960

gacagcctga ttgcaaaatt caaagaagcg ggcggaacgg tcagagagat tgatgtatga 1020

gcagagtcac cgcgattatc tccgctctgg ttatctgcat catcgtctgc ctgtcatggg 1080

ctgttaatca ttaccgtgat aacgccatta cctacaaagc ccagcgcgac aaaaatgcca 1140

gagaactgaa gctggcgaac gcggcaatta ctgacatgca gatgcgtcag cgtgatgttg 1200

ctgcgctcga tgcaaaatac acgaaggagt tagctgatgc taaagctgaa aatgatgctc 1260

tgcgtgatga tgttgccgct ggtcgtcgtc ggttgcacat caaagcagtc tgtcagtcag 1320

tgcgtgaagc caccaccgcc tccggcgtgg ataatgcagc ctccccccga ctggcagaca 1380

ccgctgaacg ggattatttc accctcagag agaggctgat cactatgcaa aaacaactgg 1440

aaggaaccca gaagtatatt aatgagcagt gcagatagag ttgcccatat cgatgggcaa 1500

ctcatgcaat tattgtgagc aatacacacg cgcttccagc ggagtataaa tgcctaaagt 1560

aataaaaccg agcaatccat ttacgaatgt ttgctgggtt tctgttttaa caacattttc 1620

-continued

tgcgccgcca caaattttgg ctgcatcgac agttttcttc tgcccaattc cagaaacgaa 1680

gaaatgatgg gtgatggttt cctttggtgc tactgctgcc ggtttgtttt gaacagtaaa 1740

cgtctgttga gcacatcctg taataagcag ggccagcgca gtagcgagta gcatttttt 1800

catggtgtta ttcccgatgc tttttgaagt tcgcagaatc gtatgtgtag aaaattaaac 1860

aaaccctaaa caatgagttg aaatttcata ttgttaatat ttattaatgt atgtcaggtg 1920

cgatgaatcg tcattgtatt cccggattaa ctatgtccac agccctgacg gggaacttct 1980

ctgcgggagt gtccgggaat aattaaaacg atgcacacag ggtttagcgc gtacacgtat 2040

tgcattatgc caacgccccg gtgctgacac ggaagaaacc ggacgttatg atttagcgtg 2100

gaaagatttg tgtagtgttc tgaatgctct cagtaaatag taatgaatta tcaaaggtat 2160

agtaatatct tttatgttca tggatatttg taacccatcg gaaaactcct gctttagcaa 2220

gattttccct gtattgctga aatgtgattt ctcttgattt caacctatca taggacgttt 2280

ctataagatg cgtgtttctt gagaatttaa catttacaac ctttttaagt ccttttatta 2340

acacggtgtt atcgttttct aacacgatgt gaatattatc tgtggctaga tagtaaatat 2400

aatgtgagac gttgtgacgt tttagttcag aataaaacaa ttcacagtct aaatcttttc 2460

gcacttgatc gaatatttct ttaaaaatgg caacctgagc cattggtaaa accttccatg 2520

tgatacgagg gcgcgtagtt tgcattatcg tttttatcgt ttcaatctgg tctgacctcc 2580

ttgtgttttg ttgatgattt atgtcaaata ttaggaatgt tttcacttaa tagtattggt 2640

tgcgtaacaa agtgcggtcc tgctggcatt ctggagggaa atacaaccga cagatgtatg 2700

taaggccaac gtgctcaaat cttcatacag aaagatttga agtaatattt taaccgctag 2760

atgaagagca agcgcatgga gcgacaaaat gaataaagaa caatctgctg atgatccctc 2820

cgtggatctg attcgtgtaa aaaatatgct taatagcacc atttctatga gttaccctga 2880

tgttgtaatt gcatgtatag aacataaggt gtctctggaa gcattcagag caattgaggc 2940

agcgttggtg aagcacgata ataatatgaa ggattattcc ctggtggttg actgatcacc 3000

ataactgcta atcattcaaa ctatttagtc tgtgacagag ccaacacgca gtctgtcact 3060

gtcaggaaag tggtaaaact gcaactcaat tactgcaatg ccctcgtaat taagtgaatt 3120

tacaatatcg tcctgttcgg agggaagaac gcgggatgtt cattcttcat cacttttaat 3180

tgatgtatat gctctctttt ctgacgttag tctccgacgg caggcttcaa tgacccaggc 3240

tgagaaattc ccggaccctt tttgctcaag agcgatgtta atttgttcaa tcatttggtt 3300

aggaaagcgg atgttgcggg ttgttgttct gcgggttctg ttcttcgttg acatgaggtt 3360

gccccgtatt cagtgtcgct gatttgtatt gtctgaagtt gtttttacgt taagttgatg 3420

cagatcaatt aatacgatac ctgcgtcata attgattatt tgacgtggtt tgatggcctc 3480

cacgcacgtt gtgatatgta gatgataatc attatcactt tacgggtcct ttccggtgaa 3540

aaaaaaggta ccaaaaaaaa catcgtcgtg agtagtgaac cgtaagca 3588

<210> SEQ ID NO 29
<211> LENGTH: 28
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic Polynucleotide

-continued

<400> SEQUENCE: 29

gcaatatcag caccaacaga aacaacct 28

1. A mutant Msp monomer comprising a variant of the sequence shown in SEQ ID NO: 2, wherein the variant:

(a) does not comprise aspartic acid (D) at position 90;

(b) does not comprise aspartic acid (D) at position 91;

(c) comprises aspartic acid (D) or glutamic acid (E) at position 93; and

(d) comprises one or more modifications which decrease the net negative charge of the inward facing amino acids in the cap forming region and/or the barrel forming region of the monomer.

2. A mutant monomer according to claim 1, wherein the cap forming region comprises amino acids 1 to 72 and 122 to 184 of SEQ ID NO: 2.

3. A mutant monomer according to claim 1, wherein the barrel forming region comprises amino acids 73 to 82 and 112 to 121 of SEQ ID NO: 2.

4. A mutant monomer according to claim 1, wherein the inward facing amino acids in the cap forming region are V9, Q12, D13, R14, T15, W40, I49, P53, G54, D56, E57, E59, T61, E63, Y66, Q67, I68, F70, P123, I125, Q126, E127, V128, A129, T130, F131, S132, V133, D134, S136, G137, E139, V144, H148, T150, V151, T152, F163, R165, I167, S169, T170 and S173.

5. A mutant monomer according to claim 1, wherein the inward facing amino acids in the barrel forming region are S73, G75, G77, N79, S81, G112, S114, S116, D118 and G120.

6. A mutant monomer according to claim 1, wherein the one or more modifications are one or more deletions of negatively charged amino acids or one or more substitutions of negatively charged amino acids with one or more positively charged, uncharged, non-polar and/or aromatic amino acids.

7. A mutant monomer according to claim 6, wherein the one or more negatively charged amino acids are substituted with alanine (A), valine (V), asparagine (N) or glycine (G).

8. A mutant monomer according to claim 1, wherein the one or more modifications are one or more introductions of positively charged amino acids.

9. A mutant monomer according to claim 6, wherein the one or more positively charged amino acids are histidine (H), lysine (K) and/or arginine (R).

10. A mutant monomer according to claim 1, wherein the one or more modifications are one or more chemical modifications of one or more negatively charged amino acids which neutralise their negative charge.

11. A mutant monomer according to claim 1, wherein the one or more modifications reduce the net negative charge at one or more of positions 118, 126, 134 and 139.

12. A mutant monomer according to claim 1, wherein

(i) the variant comprises a positively charged amino acid at one or more of positions 114, 116, 120, 123, 70, 73, 75, 77 and 79,

(ii) the variant comprises a positively charged amino acid at one or more of positions 123, 125, 127 and 128;

(iii) the variant comprises a positively charged amino acid at one or more of positions 129, 132, 136, 137, 59, 61 and 63;

(iv) the variant comprises a positively charged amino acid at one or more of positions 137, 138, 141, 143, 45, 47, 49 and 51;

(v) the variant does not comprise aspartic acid (D) or glutamic acid (E) at one or more of positions 118, 126, 134 and 139;

(vi) the variant comprises arginine (R), glycine (G) or asparagine (N) at one or more of positions 118, 126, 134 and 139;

(vii) the variant comprises D118R, Q126R, D134R and E139K;

(viii) the variant comprises serine (S), glutamine (Q), leucine (L), methionine (M), isoleucine (I), alanine (A), valine (V), glycine (G), phenylalanine (F), tryptophan (W), tyrosine (Y), histidine (H), threonine (T), arginine (R), lysine (K), asparagine (N) or cysteine (C) at position 90 and/or position 91;

(ix) the variant comprises asparagine (N) at position 90 and/or position 91;

(x) the variant comprises one or more of:

(e) serine (S) at position 75;

(f) serine (S) at position 77; and

(g) asparagine (N) or lysine (K) at position 88;

(xi) the variant comprises G75S, G77S and L88K or G75S, G77S and L88N;

(xii) the variant comprises G75S, G77S, L88N, D90N, D91N, D118R, Q126R, D134R and E139K; and/or

(xiii) the variant further comprises one or more of:

(h) phenylalanine (F) at position 89;

(i) glutamic acid (E) at position 95 and lysine (K) at position 98;

(j) aspartic acid (D) at position 96;

(k) glycine (G) at position 102;

(l) alanine (A) at position 103; and

(m) alanine (A), serine (S) or proline (P) at position 108.

13-32. (canceled)

33. A pore derived from Msp comprising at least one mutant monomer according to claim 1.

34. A pore according to claim 33, wherein the pore comprises eight mutant monomers according to claim 1 and wherein at least one of them differs from the others.

35. A method of characterising a target polynucleotide, comprising:

a) contacting the polynucleotide with a pore according to claim 31 such that the polynucleotide moves through the pore; and

b) taking one or more measurements as the polynucleotide moves with respect to the pore, wherein the measurements are indicative of one or more characteristics of the polynucleotide, and thereby characterising the target polynucleotide.

36. A method according to claim 35, wherein the one or more characteristics are selected from (i) the length of the

polynucleotide, (ii) the identity of the polynucleotide, (iii) the sequence of the polynucleotide, (iv) the secondary structure of the polynucleotide and (v) whether or not the polynucleotide is modified.

37. A method according to claim **35**, wherein the one or more characteristics of the polynucleotide are measured by electrical measurement and/or optical measurement.

38. (canceled)

39. A method according to claim **35**, wherein step a) further comprises contacting the polynucleotide with a polynucleotide binding protein such that the protein controls the movement of the polynucleotide through the pore.

40. (canceled)

41. A method according to claim **39**, wherein the polynucleotide binding protein is a helicase or is derived from a helicase.

42-51. (canceled)

52. A pore according to claim **33**, wherein the pore comprises eight identical mutant monomers according to claim **1**.

* * * * *