

Claims

1. A method of assessing a biological sequence capable of a preselected function, comprising:
 - 5 (a) preselecting a biological molecule, wherein the biological molecule is capable of a function of interest;
 - (b) preparing a testing sequence database comprising a plurality of sequence fragments of the preselected biological molecule, wherein the preselected sequence fragments are a predetermined length;
 - 10 (c) fragmenting the sequence of one or more test biological molecules into lengths equivalent to the predetermined length of the sequence fragments of the preselected biological molecule in the testing sequence database;
 - (d) detecting a presence or absence of a sequence match between the sequence of at least one fragment of the fragmented test biological molecules and at least one of the plurality of sequence fragments of the preselected biological sequence, and
 - 15 (e) acting in response to the detection in (d), wherein the detecting in (d) provides an assessment of the test biological molecule.
2. The method of claim 1, wherein the acting in response to (d) comprises one of more of: preventing synthesis of the test biological molecule, permitting synthesis of the test biological
20 molecule, sequencing one or more polynucleotide molecules, DNA sequencing, DNA molecule design, polypeptide sequence determination, and further sequence identification steps.
3. The method of claim 1, further comprising identifying in the testing sequence database one or more sequence fragments of the preselected biological sequence that match one or more
25 sequence fragments, respectively, of a second biological molecule having a biological function unrelated to the biological function of interest of the preselected biological molecule, and removing the identified sequence fragments(s) from the testing sequence database.
4. The method of claim 1, wherein if the presence of a sequence match is detected in (d) the
30 action comprises preventing synthesis of the test biological molecule.

5. The method of any one of claims 1, wherein a means of preparing the testing sequence database comprises:
- (a) screening the plurality of sequence fragments of the preselected biological sequence molecule against at least one control sequence database, wherein the control sequence database comprises a plurality of control sequence fragments of at least one molecule capable of a function of interest unrelated to the function of interest of the preselected biological molecule;
 - (b) identifying the presence of a match between a sequence fragment in the plurality of sequence fragments of the preselected biological molecule and a sequence fragment in the control sequence database that is a fragment of the biological molecule identified as capable of a function unrelated to the function of interest of the preselected biological molecule; and
 - (c) removing from the testing sequence database the sequence fragment of the preselected biological sequence identified as matching the sequence fragment of the biological sequence identified as capable of a function of interest unrelated to the function of interest of the molecule capable of the function of interest.
6. The method of any one of claim 1, wherein the preselected biological molecule is a polynucleotide.
7. The method of claim 6, wherein the sequence of the biological molecule is a full-length nucleic acid sequence of the polynucleotide or is a portion of the full-length nucleic acid sequence of the polynucleotide.
8. The method of claim 7, wherein the full-length nucleic acid sequence encodes a protein.
9. The method of any one of claims 6-8, wherein the predetermined length of the sequence fragments of the preselected polynucleotide molecule is 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, or more nucleotides.
10. The method of any one of claims 1-6, wherein the preselected biological molecule comprises a polypeptide.

11. The method of claim 10, wherein the amino acid sequence of the preselected biological molecule is a full-length amino acid sequence of the polypeptide, or is a portion of the full-length amino acid sequence of the polypeptide.

5

12. The method of claim 10, wherein the predetermined length of the sequence fragments of the preselected polypeptide molecule is 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 91, 2, 93, 94, 95, or more amino acids.

10

13. The method of any claim 1, wherein the plurality of sequence fragments of the preselected biological molecule: (1) comprises all or a significant portion of possible fragments or at least one essential fragment of the biological molecule that is capable of the function of interest, and (2) does not comprise sequences found in a biological molecule capable of a function unrelated to the function of the preselected biological molecule.

15

14. The method of claim 2, wherein the control sequence database comprises a plurality of control sequence fragments of at least one molecule capable of a function of interest unrelated to the function of interest of the preselected biological molecule.

20

15. The method of claim 2, wherein the control sequence database comprises a plurality of control sequence fragments of at least one molecule not capable of the function of interest of the preselected biological molecule.

25

16. The method of claim 1, wherein the predetermined length is the same for all fragments of the preselected biological molecule.

17. The method of claim 1, wherein the predetermined length of the sequence fragments of the preselected biological molecule comprises more than one length.

30

18. The method of claim 1, wherein the testing sequence database comprises one or more sequence fragments randomly or pseudorandomly selected from sequences of molecules known to be capable of a function different from the preselected molecule's function of interest.
- 5 19. The method of claim 18, wherein the randomly or pseudorandomly selected sequence fragments are biased towards sequence regions with greater homology to functionally or phylogenetically related sequences.
20. The method of claim 1, wherein the testing sequence database further comprises
10 sequences that are functional equivalents of the plurality of sequence fragments of the preselected biological molecule.
21. The method of claim 20, wherein a means for identifying the functional equivalents comprises a computational means.
- 15 22. The method of claim 20, wherein a means for identifying the functional equivalents comprises an experimental means.
23. The method of any one of claims 20-22, wherein a computational means for selecting the
20 functional equivalents included in the testing sequence database comprises using a classifier based on experimental data to evaluate the accuracy of the computational means.
24. The method of claim 20, wherein a means for selecting the functional equivalents included in the testing sequence database comprises inclusion of a minimal number of sequences
25 calculated to achieve a predetermined likelihood of successfully preventing a test sequence from escaping detection.
25. The method of any one of claims 20-22, wherein a means for selecting the functional
30 equivalents included in the testing sequence database comprises a random selection method or a pseudorandom selection method.

26. The method of claim 1, wherein the identities of all sequence fragments of one or both of the testing sequence database and the test biological molecule are protected.
27. The method of claim 26, wherein a means of the protecting comprises application of a cryptographic hash function, wherein the cryptographic hash function deterministically maps the sequence data to a bit string of fixed size using a one-way function.
28. The method of claim 27, wherein the application of the cryptographic hash function cannot be reversed without a brute-force search of all possible sequence inputs into the testing sequence database.
29. The method of claim 28, wherein the application of the cryptographic hash function further comprises use of one or more information keys that must be accessed to attempt the brute-force search.
30. The method of claim 29, wherein the application of the cryptographic hash function requires keys from a plurality of independent sources that must cooperate to compute the hash without any one server gaining access to the sequence data.
31. The method of claim 30, wherein the independent sources comprise independent computer servers.
32. The method of claim 1, further comprising dividing the prepared testing sequence database into two or more partial testing sequence databases, and the prepared testing sequence database used for detecting of the presence or absence of a sequence match is one two or more partial testing sequences databases.
33. The method of claim 32, wherein if a sequence match is detected between the partial testing sequence database and one or more fragments of the test biological molecules, the method further comprises detecting the presence or absence of a sequence using another of the two or more partial testing sequence databases.

34. The method of claim 1, wherein the testing sequence database contains a portion of a larger database of sequence fragments such that the fragments included in the testing sequence database can be rotated frequently or upon a match being discovered.

5

35. A method of identifying a biological sequence capable of a preselected function, comprising:

- (a) preselecting a biological molecule, wherein the preselected biological molecule is capable of a function of interest;
- 10 (b) preparing a testing sequence database comprising a plurality of sequence fragments of the preselected biological molecule, wherein the preselected sequence fragments are a predetermined length;
- (c) fragmenting the sequence of one or more test biological molecules into lengths equivalent to the predetermined length of the sequence fragments the preselected biological
- 15 molecule in the testing sequence database; and
- (d) detecting a presence or absence of a sequence match between the sequence of at least one fragment of the fragmented test biological molecules and at least one of the plurality of sequence fragments of the preselected biological sequence;
wherein a means of preparing the testing sequence database comprises:
 - 20 (i) screening the plurality of sequence fragments of the preselected biological sequence molecule against at least one control sequence database, wherein the control sequence database comprises a plurality of control sequence fragments of at least one molecule capable of a function of interest unrelated to the function of interest of the preselected biological molecule;
 - 25 (ii) identifying the presence of a match between a sequence fragment in the plurality of sequence fragments of the preselected biological molecule and a sequence fragment in the control sequence database that is a fragment of the biological molecule identified as capable of a function unrelated to the function of interest of the preselected biological molecule; and
 - (iii) removing from the testing sequence database the sequence fragment of the
 - 30 preselected biological sequence identified as matching the sequence fragment of the

biological sequence identified as capable of a function of interest unrelated to the function of interest of the molecule capable of the function of interest.

36. The method of 35, wherein the preselected biological molecule is a polynucleotide.
- 5 37. The method of claim 36, wherein the sequence of the preselected biological molecule is a full-length nucleic acid sequence of the polynucleotide or is a portion of the full-length nucleic acid sequence of the polynucleotide.
- 10 38. The method of claim 37, wherein the full-length nucleic acid encodes a protein.
39. The method of claim 36, wherein the predetermined length of the sequence fragments of the preselected polynucleotide molecule is 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 50, 55, 60, 65, 70, 75, 80, 85, 90, 15 95, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, or more nucleotides.
40. The method of claim 35, wherein the preselected biological molecule comprises a polypeptide molecule.
- 20 41. The method of claim 40, wherein the sequence of the preselected biological molecule is a full-length amino acid sequence of the polypeptide, or is a portion of the full-length amino acid sequence of the polypeptide.
42. The method of claim 40 or 41, wherein the predetermined length of the sequence 25 fragments of the preselected polynucleotide molecule is 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 91, 2, 93, 94, 95, or more amino acids.
43. The method of claim 35, wherein the plurality of sequence fragments of the preselected 30 biological molecule: (1) comprises all or a significant portion of possible fragments of the biological molecule that is capable of the function of interest, and (2) does not comprise

sequences not found in a biological molecule capable of a function unrelated to the function of the preselected biological molecule.

44. The method of claim 35, wherein the control sequence database comprises a plurality of control sequence fragments of at least one molecule capable of a function of interest unrelated to the function of interest of the preselected biological molecule.

45. The method of claim 35, wherein the control sequence database comprises a plurality of control sequence fragments of at least one molecule not capable of the function of interest of the preselected biological molecule.

46. The method of claim 35, wherein a single length is the predetermined length of the sequence fragments of the preselected biological molecule.

47. The method of claim 35, wherein there is more than one predetermined length of the sequence fragments of the preselected biological molecule.

48. The method of claim 35, wherein the testing sequence database comprises one or more sequence fragments randomly selected from sequences of molecules known to be capable of a function different from the preselected molecule's function of interest.

49. The method of claim 48, wherein the randomly selected sequence fragments are biased towards conserved regions.

50. The method of claim 35, wherein the testing database further comprises sequences that are functional equivalents of the plurality of sequence fragments of the preselected biological molecule.

51. The method of claim 35, wherein the identities of all sequence fragments of one or both of the testing sequence database and the test biological molecule are protected.

52. The method of claim 51, wherein a means of the protecting comprises application of a cryptographic hash function, wherein the cryptographic hash function deterministically maps the sequence data to a bit string of fixed size using a one-way function.
- 5 53. The method of claim 52, wherein the application of the cryptographic hash function cannot be reversed without a brute-force search of all possible sequence inputs into the testing sequence database.
54. The method of claim 53, wherein the application of the cryptographic hash function
10 further comprises use of one or more information keys that must be accessed to attempt the brute-force search.
55. The method of claim 51, wherein the application of the cryptographic hash function requires keys from a plurality of independent sources that must cooperate to compute the hash
15 without any one server gaining access to the sequence data.
56. The method of claim 55, wherein the independent sources comprise independent computer servers.
- 20 57. The method of claim 35, further comprising dividing the prepared testing sequence database into two or more partial testing sequence databases, and the prepared testing sequence database used for detecting of the presence or absence of a sequence match is one two or more partial testing sequences databases.
- 25 58. The method of claim 57, wherein if a sequence match is detected between the partial testing sequence database and one or more fragments of the test biological molecules, the method further comprises detecting the presence or absence of a sequence using another of the two or more partial testing sequence databases.

59. The method of claim 35, wherein the testing sequence database contains a portion of a larger database of sequence fragments such that the fragments included in the testing sequence database can be rotated frequently or upon a match being discovered.

5 60. The method of claim 35, further comprising acting in response to the detecting in claim 35(d), wherein the acting comprises one of more of: preventing synthesis of the test biological molecule, permitting synthesis of the test biological molecule, sequencing one or more polynucleotide molecules, DNA sequencing, DNA molecule design, determining an amino acid sequence of a polypeptide, and further sequence identification steps.

10

61. The method of claim 35, wherein if the presence of a sequence match is detected in claim 35(d) the action comprises preventing synthesis of the test biological molecule.

15 62. The method of claim 35, further comprising identifying in the prepared testing sequence database one or more sequence fragments of the preselected biological sequence that match one or more sequence fragments, respectively, of a second biological molecule having a biological function unrelated to the biological function of interest of the preselected biological molecule, and removing the identified sequence fragments(s) from the testing sequence database

20 63. A testing sequence database prepared by a method of any one of claims 1-34.

64. A testing sequence database prepared by a method of claim 1.

25 65. A method of assessing a biological sequence using a testing sequence database of claim 63.

66. The method of claim 65, wherein the assessing comprises determining whether to permit or prevent synthesis of the assessed biological sequence.

30 67. A testing sequence database prepared with a method of any one of claims 35- 62.

68. A testing sequence database prepared with a method of claim 35.

69. A method of assessing a biological sequence using a testing sequence database of claim 67.

5

70. The method of claim 69, wherein the assessing comprises determining whether to permit or prevent synthesis of the assessed biological sequence.