SUMMARY

The present invention provides methods for identifying bio-molecules with desired properties, or which are most suitable for acquiring such properties, from complex bio-molecule libraries or sets of such libraries. More specifically, some embodiments of the present invention provide methods for building sequence-activity models that include a product of a plurality of multiplicative terms and using the models to guide directed evolution.

According to an aspect of the present invention, there is provided a method of conducting directed evolution, the method comprising, (a) obtaining sequence and activity data for each of a plurality of protein variants; (b) generating a sequence-activity model from the sequence and activity data for each of the plurality of protein variants, wherein the sequence-activity model comprises: a product of multiple terms, wherein each of at least some of the terms comprises a coefficient representing the contribution to activity of a defined amino acid or nucleotide at a defined position in a protein or nucleic acid sequence, and a dependent variable representing the activity of the protein variants; (c) selecting one or more mutations for a round of directed evolution by evaluating the coefficients of the terms of the sequence-activity model to identify one or more of the defined amino acids or nucleotides at the defined positions that contribute to the activity; (d) preparing a plurality of oligonucleotides containing or encoding the one or more mutations selected in (c); and (e) performing a round of directed evolution using the plurality of oligonucleotides prepared in (d).

According to another aspect of the present invention, there is provided a method of conducting directed evolution, the method comprising: (a) obtaining sequence and activity data for each of a plurality of protein variants; (b) generating a sequence-activity model from the sequence and activity data for each of the plurality of protein variants, wherein the sequence-activity model comprises: a product of multiple terms, wherein each of at least some of the terms comprises a coefficient representing the contribution to activity of a defined amino acid or nucleotide at a defined position in a protein sequence or a nucleic acid sequence encoding the protein sequence; and a dependent variable representing the activity of the protein variants; (c) selecting one or more mutations by evaluating the coefficients of the terms of the sequence-activity model to identify one or more of the defined amino acids or nucleotides at the defined positions that contribute to the activity; (d) identifying a new protein or a new nucleic acid sequence comprising the one or more mutations

selected in (c); and (e) preparing and assaying the new protein or a protein encoded by the new nucleic acid sequence.

According to another aspect of the present invention, there is provided a method of conducting directed evolution, the method comprising: (a) obtaining sequence and activity data for each of a plurality of protein variants; (b) generating a sequence-activity model from the sequence and activity data for each of the plurality of protein variants, wherein the sequence-activity model comprises: a product of multiple terms, wherein each of at least some of the terms comprises a coefficient representing the contribution to activity of a defined amino acid or nucleotide at a defined position in a protein sequence or a nucleic acid sequence encoding the protein sequence; and a dependent variable representing the activity of the protein variants; (c) selecting one or more positions in the protein sequence or nucleic acid sequence by evaluating the coefficients of the terms of the sequence-activity model to identify one or more of the defined amino acids or nucleotides at the defined positions that contribute to the activity; and (d) conducting saturation mutagenesis at the one or more positions identified in (c).

According to another aspect of the present invention, there is provided a method of conducting directed evolution, the method comprising: (a) obtaining sequence and activity data for each of a plurality of protein variants; (b) generating a sequence-activity model from the sequence and activity data for each of the plurality of protein variants, wherein the sequence-activity model comprises: a product of multiple terms, wherein each of at least some of the terms comprises a coefficient representing the contribution to activity of a defined amino acid or nucleotide at a defined position in a protein sequence or a nucleic acid sequence encoding the protein sequence; and a dependent variable representing the activity of the protein variants; (c) applying multiple protein sequences or multiple amino acid sequences to the sequence-activity model and determining activity values predicted by the sequence-activity model for each of the multiple protein sequences or nucleic acid sequences; (d) selecting a new protein sequence or a new nucleic acid sequence from among the multiple protein sequences or multiple amino acid sequences applied in (c) by evaluating the activity values predicted by the sequence-activity model for the multiple sequences; and (e) preparing and assaying a protein having the new protein sequence or a protein encoded by the new nucleic acid sequence.

According to another aspect of the present invention, there is provided a method of conducting directed evolution of biological molecules, the method comprising, (a) obtaining sequence and activity data for a plurality of biological molecules, each biological molecule comprising a sequence having subunits of various types and sequence positions; (b) building a sequence-activity model from the obtained data, wherein the sequence-activity model predicts activity as a function of the type and sequence position of subunits, the sequence-activity model comprises a product of a plurality of base multiplicative terms, each of the base multiplicative terms comprising (1) a dummy variable representing the presence/absence of a defined subunit of a specific type at a specific sequence position, and (2) a coefficient representing the defined subunit's contribution to activity; and (c) using the sequence-activity model to identify one or more subunits of specific types at specific positions for variation to impact a desired activity of the biological molecules.

According to another aspect of the present invention, there is provided a computer system, comprising: one or more processors; system memory; and one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed by the one or more processors, cause the computer system to implement a method for conducting directed evolution of biological molecules, the method comprising: (a) obtaining sequence and activity data for a plurality of biological molecules, each biological molecule comprising a sequence having subunits of various types and sequence positions; (b) building a sequence-activity model from the obtained data, wherein the sequence-activity model predicts activity as a function of the type and sequence position of subunits, the sequence-activity model comprises a product of a plurality of base multiplicative terms, each of the base multiplicative terms comprising (1) a dummy variable representing the presence/absence of a defined subunit of a specific type at a specific sequence position, and (2) a coefficient representing the defined subunit's contribution to activity; and (c) using the sequence-activity model to identify one or more subunits of specific types at specific positions for variation to impact a desired activity of the biological molecules.

FORM OF THE MODEL

Some embodiments disclosed herein provide methods for building sequence-activity models. Each of the models includes a product of a plurality of multiplicative terms. At least some of the multiplicative terms are non-interaction multiplicative terms, each of which comprises a coefficient representing a defined amino acid or nucleotide's contribution to activity of interest. The defined amino acid or nucleotide is of a specific residue type at a specific position in a protein or nucleic acid sequence. Each of the non-interactive multiplicative terms also includes an independent variable (typically only one independent variable), which may take the form of a dummy variable. The independent variable represents the presence or absence of the defined amino acid or nucleotide, of a specific type at a specific sequence location. These multiplicative terms are termed as non-interaction terms because each of them represents the contribution of a single residue at single position in the protein or nucleic acid sequence. The independent variables in non-interaction terms do not represent two or more interacting residues. In addition, each of the sequence-activity models in some embodiments include a dependent variable representing the activity of a protein variant, where the model describes a relationship between the activity of the protein variant and the product of the plurality of multiplicative terms.

In some embodiments of the invention, the sequence-activity models also include interaction terms, each of which comprises an interaction coefficient representing the contribution to activity of a defined combination of (i) a first amino acid or nucleotide at a first position in the protein sequence, and (ii) a second amino acid or nucleotide at a second position in the protein sequence. The contribution represented by the interaction coefficient is independent and different from the contribution by each of the first residue and second residue in isolation.