(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 December 2018 (13.12.2018)

WIPO | PCT

(10) International Publication Number
WO 2018/227167 A1

(51) International Patent Classification:
G06F 19/18 (2011.01) *C40B 30/02* (2006.01)
G06F 19/24 (2011.01) *C40B 30/10* (2006.01)

(21) International Application Number:
PCT/US2018/036777

(22) International Filing Date:
08 June 2018 (08.06.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/517,048 08 June 2017 (08.06.2017) US

(71) Applicant: **JUST BIOTHERAPEUTICS, INC.** [US/US]; 400 Terry Avenue North, Seattle, WA 98109 (US).

(72) Inventors: **SHAVER, Jeremy, Martin**; c/o Just Biotherapeutics, Inc., 400 Terry Avenue North, Seattle, WA 98109 (US). **KETCHEM, Randal, Robert**; c/o Just Biotherapeutics, Inc., 400 Terry Avenue North, Seattle, WA 98109 (US).

(74) Agent: **PIRIO, Maurice** et al.; Perkins Coie LLP, P.O. Box 1247, Seattle, WA 98111-1247 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ,

WO 2018/227167 A1

(54) Title: PREDICTING MOLECULAR PROPERTIES OF MOLECULAR VARIANTS USING RESIDUE-SPECIFIC MOLECULAR STRUCTURAL FEATURES

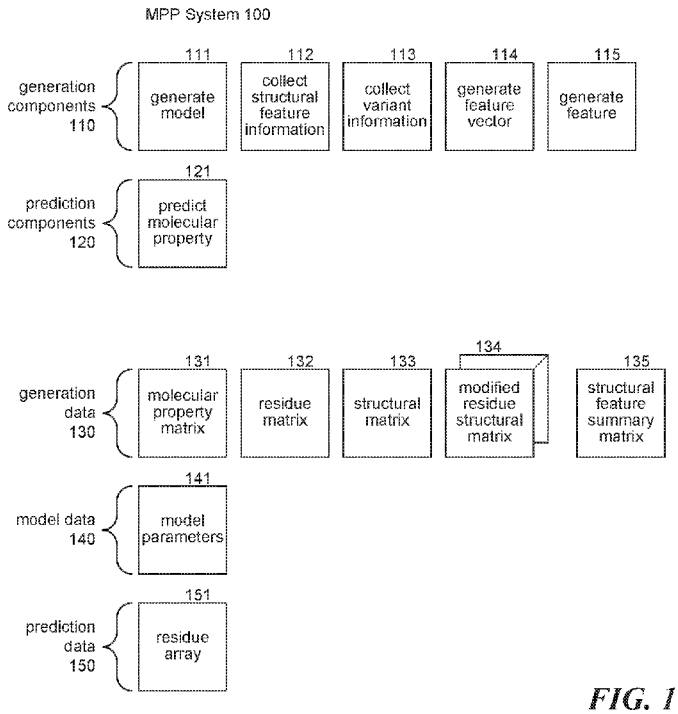

FIG. 1

(57) Abstract: A system for generating a model for predicting a molecular property of a variant of a molecule is provided. For each of a plurality of variants of the molecule, the system for each structural feature, aggregates the values for the structural features of the residues of the molecule that were modified to form the variant to form a feature vector for the variant. The system assigns the value for the molecular property of the variant to the feature vector wherein the feature vector and the assigned value form training data. The system then generates the model for predicting a value for the molecular property using the training data for the plurality of variants.

[Continued on next page]

UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

PREDICTING MOLECULAR PROPERTIES OF MOLECULAR VARIANTS USING RESIDUE-SPECIFIC MOLECULAR STRUCTURAL FEATURES

BACKGROUND

[0001] Because the cost of developing breakthrough therapeutics based on modern biotechnology is so high, such therapeutics are not available to most people. A contributing factor to the high cost is that it is difficult to identify the molecular properties of new variants of a molecule such as an antibody. Although various tools are available to help in determining the molecular properties of molecules, they rely in large part in being able to synthesize the variant, which itself can be costly and time consuming.

BRIEF DESCRIPTION OF THE DRAWINGS

[0002] Figure 1 is a block diagram illustrating components of the MPP system in some embodiments.

[0003] Figure 2 is a flow diagram that illustrates the processing of a generate model component of the MPP system in some embodiments.

[0004] Figure 3 is a flow diagram that illustrates the processing of a collect structural feature information component of the MPP system in some embodiments.

[0005] Figure 4 is a flow diagram that illustrates the processing of a collect variant information component of the MPP system in some embodiments.

[0006] Figure 5 is a flow diagram that illustrates the processing of a generate feature vector component of the MPP system in some embodiments.

[0007] Figure 6 is a flow diagram that illustrates the processing of a generate feature component of the MPP system in some embodiments.

[0008] Figure 7 is a flow diagram that illustrates the processing of a predict molecular property component of the MPP system in some embodiments.

DETAILED DESCRIPTION

[0009] A method and system for estimating molecular properties of new variants of a parent molecule (e.g., an antibody molecule) prior to synthesis of the new variants is provided. In some embodiments, a molecular property prediction ("MPP") system uses of various structural features of residues in a parent molecule (referred to as a molecule) in conjunction with molecular properties measured for a set of variants of the molecule. The MPP system supports predicting molecular properties of a new variant without having to calculate structural models for each of the variants. The MPP system also avoids creating the much more complicated "all molecules" model which attempts to predict molecular properties for any general antibody. The MPP system provides a prediction model that is more specific to a given molecule and more robust than the "all molecules" model. In some embodiments, the MPP system predicts molecular properties that include, but are not limited to, molecular characterizations such as antibody melting temperature ("Tm"), the percentage of high molecular weight expected during the expression and purification of the variant ("HMW"), chemical unfolding behavior, solubility, viscosity, and aggregation behavior (e.g., self-interaction nanoparticle spectroscopy--"SINS").

[0010] In some embodiments, the MPP system generates a model for predicting a molecular property of a variant of a molecule. The MPP system accesses values for structural features of the residues of the molecule. For example, the structural features of a molecule may include participation of the residue in charge patches or hydrophobic patches and group diversity of neighboring residues. For each variant of the molecule, the MPP system accesses variant information indicating which residues in a sequence of residues of the molecule were modified to form the variant and a value for the molecular property of the variant. For each structural feature, the MPP system aggregates the values for the structural features of the residues of the molecule that were modified to form the variant to generate a feature vector for the variant. The MPP system assigns the value for the molecular property of the variant to the feature vector. The feature vector and the assigned value for each variant form training data. The MPP system then uses the training data to generate the prediction model for predicting a value for the molecular property. For example, the MPP system may use a linear regression technique, neural networks,

random forest techniques, Lasso regression techniques, and partial least square regression techniques to generate the prediction model. The MPP system may generate a separate prediction model for each molecular property.

[0011] After the prediction model is generated, the MPP system can then use the prediction model to predict values for a molecular property for a new variant of the molecule without having to synthesize the new variant and provide information to help guide future variant creation and experimentations. The MPP system inputs an indication of the residues of the molecule that are to be changed (e.g., by substituting one amino acid for another). The MPP system generates a new feature vector for the variant in a manner similar to how the feature vectors of the training data are generated. The MPP system then applies the prediction model for a molecular property to the new feature vector to predict the value for the molecular property.

[0012] Figure 1 is a block diagram illustrating components of the MPP system in some embodiments. The MPP system 100 includes generation components 110 prediction components 120, generation data 130, model data 140, and prediction data 150. The generation components include a generate model component 111, a collect structural feature information component 112, a collect variant information component 113, a generate feature vector component 114, and a generate feature component 115. The generate model component 111 controls the overall generation of the model by invoking the collect structural feature information component 112, the collect variant information component 113, and the generate feature vector component 114 and then training the model. The collect structural feature information component 112 collects the values for the structural features for the residues of the molecule. The structural features information is stored in a structural matrix 133 that includes a row for each residue and a column for each feature with entries indicating the values of a feature for a residue of the molecule. The collect variant information component 113 collects information on the variants of the molecules that includes which residues of the molecule were modified and a value for each molecular property for the variant. The information relating to the modified variants is stored in a residue matrix 132 that includes a row for each variant and a column for each residue with the entries indicating whether that residue is modified in that variant. The

information related to the values for the molecular property are stored in a molecular property matrix 131 that includes a row for each variant and a column for each molecular property with the entries indicating the values for each molecular property for each variant. The generate feature vector component 114 generates a feature vector of features for each variant. The feature vector for a variant includes, for each structural feature, one or more statistics generated from the values for that structural feature of the residues of that variant that were modified. To generate a feature vector for a variant, the generate feature vector component 114 may generate a modified residue structural matrix 134 for that variant that includes a column for each modified residue of that variant and a row for each feature with entries indicating the values for that structural feature in that residue for the molecule. The generate feature component 115 then aggregates the values from the modified residue structural matrix 134 for a variant. The generate feature component 115 may generate a structural feature summary matrix 135 that includes a row for each variant and a column for each feature with entries indicating the value for the feature that may be a statistic such as sum, mean, and standard deviation for that variant. Each row of the structural feature summary matrix 135 represents the feature vector for a variant. The generate model component then assigns to each feature vector of a variant the value of its molecular property from the molecular property matrix 131.

[0013] The generate model component 111 then trains the prediction model using the feature vectors with their assigned values. The generate model component 111 stores the parameters learned during the training in a model parameters store 141. Once the prediction model is generated, the value of a molecular property for a new variant can be predicted using a predict molecular property component 121. The predict molecular property component inputs a residue array 151 that indicates for each residue of the molecule whether the corresponding residue in the new variant has been modified. The predict molecular property component 121 invokes the generate feature vector component to generate a new feature vector for the new variant. The predict molecular property component 121 then applies the model to the new feature vector to predict the value for the molecular property of the new variant.

[0014] The matrices below provide example values for the various matrices of the MPP system. The structural matrix is generated for variant 2. The modified residue structural matrix includes the three rows of the structural matrix corresponding to the entries of the residue matrix that have values of "true." The structural feature summary matrix includes a row for each variant with a column for each statistic--the maximum, mean, and standard deviation for each of the hydrophobic area, the positive area, and solvent accessibility ("SA") structural features. For example, the maximum, mean, and standard deviation ("SD") of the positive area for variant 2 are 180, 60, and 84.853, respectively.

Molecular Property Matrix

Molecular Property

Variant	HMW
1	0.60083
2	0.87401
3	0.53155

Residue Matrix

Residue

Variant	1	2	3	4	5
1	False	False	False	True	True
2	True	False	False	True	True
3	False	True	True	True	True

Structural Matrix

Structural Feature

Residue	hydrophobic area	positive area	SA
1	0	0	36.1
2	0	0	23.5
3	110	180	57.1
4	0	180	199.1
5	0	0	84.7

Modified Residue Structural Matrix (for Variant 2)

Structural Feature

Residue	hydrophobic area	positive area	SA
1	0	0	36.1
4	0	180	199.1
5	0	0	84.7

Structural Feature Summary Matrix

Statistic

Variant	SA			hydrophobic area			positive area		
	max	mean	SD	max	mean	SD	max	mean	SD
1	199.1	141.9	57.2	0	0	0	180	90	90
2	199.1	106.63	68.328	0	0	0	180	60	84.853
3	199.1	91.1	66.014	110	27.5	47.631	180	90	90

[0015] The computing systems on which the MPP system may be implemented may include a central processing unit, input devices, output devices (e.g., display devices and speakers), storage devices (e.g., memory and disk drives), network interfaces, graphics processing units, and so on. The input devices may include keyboards, pointing devices, touch screens, gesture recognition devices (e.g., for air gestures), head and eye tracking devices, microphones for voice recognition, and so on. The computing systems may include desktop computers, laptops, tablets, servers, and so on. The computing systems may access computer-readable media that include computer-readable storage media and data transmission media. The computer-readable storage media are tangible storage means that do not include a transitory, propagating signal. Examples of computer-readable storage media include memory such as primary memory, cache memory, and secondary memory (e.g., DVD) and other storage. The computer-readable storage media may have recorded on it or may be encoded with computer-executable instructions or logic that implements the MPP system. The data transmission media is used for transmitting data via transitory, propagating signals or carrier waves (e.g., electromagnetism) via a wired or wireless connection.

[0016] The MPP system may be described in the general context of computer-executable instructions, such as program modules and components, executed by one or more computers, processors, or other devices. Generally, program modules or components include routines, programs, objects, data structures, and so on that perform particular tasks or implement particular data types. Typically, the functionality of the program modules may be combined or distributed as desired in various examples. Aspects of the MPP system may be implemented in hardware using, for example, an application-specific integrated circuit (ASIC) or field programmable gate array ("FPGA").

[0017] In some embodiments, the actual molecular properties are measured for each previously synthesized variant (i.e., that are used for training) and are referred to as the "Y values." A separate value is collected for each variant (rows) and each molecular property (e.g. Tm, HMW, SINS – as columns) in the molecular property matrix 131.

[0018] In some embodiments, the residue matrix 132 is a Boolean matrix that describes which residue(s) were modified in the parent molecule to create the given

variant. Each row of this matrix represents one variant and each column a Boolean vector indicating if a given residue was mutated.

[0019] In some embodiments, the structural features of the molecule for each residue that is available for modification in the molecule is stored in the structural matrix 133. The structural matrix 133 contains columns representing the individual structural features and rows as the residues. The structural matrix 133 may be derived from a structural model of the molecule by extracting for each residue, value for structural features such as:

- participation of the residue in positive or negative charge patches or hydrophobic patches
- group diversity of neighboring residues (e.g. number of hydrophobic, acidic, basic, or neutral residues within a given distance – distance based on the tertiary structure)
- solvent accessibility of the residue (high at the surface of the molecule)
- nearness to region/chain interfaces (e.g. distance to Fv and constant domain interface)
- secondary structure environment
- original (in molecule) and new residue (in variant) length or size
- original and new residue pKa (i.e., acidity)

[0020] A goal of the MPP system may be to identify a subset of structural features that can be used to estimate a given molecular property for both the set of previously synthesized variants (known as a calibration set) and to estimate the molecular property for new variants (without the synthesis and measurement of the molecular property of those variants). The MPP system starts by using the residue matrix 132 to extract the rows of the structural matrix 133 for residues that were modified for a given variant. The modified residue structural matrix 134 for a variant has as many rows as the number of residues modified for that variant. The modified residue structural matrix 134 is compressed down into a single row by applying a set of statistical metrics to each column including, but not necessarily limited to: sum, mean, standard deviation, skew, kurtosis, minimum, maximum, product, and sum and mean of the log of absolute values (e.g., logs post-multiplied by original value sign either summed or averaged). The result of applying

each of these operations on each original structural feature column is that the statistic is turned into a new column. Accordingly, there are 10 new columns (e.g., given the set of 10 statistical metrics above) for each original structural feature. The MPP system thus captures the molecular variation of the set of modified residues without having to specifically list the residues. After iterating over all variants, the MPP system generates a structural feature summary matrix 135 in which each row is a variant and the columns are the statistical summary of each set of modified structural features. Next, the structural feature summary matrix 135 is either used in a dimension-reducing regression or classification model (e.g., partial least squares, or neural network with reducing hidden-layer nodes) to predict the Y values (e.g., molecular properties), or it is used in a variable selection method (e.g., a genetic algorithm, or correlation-based selection) to reduce the number of variables. In the latter case, only the selected variables are used in a regression or classification model.

[0021] In some embodiments, the MPP system may support molecule-generalized models. Because each molecule has a different starting value for a given molecular property, and because each molecule may be differently sensitive to the characteristics of the modified residues, models as described above are expected to only be applicable to the specific parent molecule. However, some property predictions can be made less molecule-specific by adding molecule-encoding variables or doing *a priori* corrections, for example, by making the Y values be relative to the parent molecule. In such cases, multiple structural feature summary matrices for different molecules may be combined along with their corresponding Y values and processed in a single model. For example, a combined structural feature summary matrix may include a column to identify the parent molecule of a variant. In such a case, the MPP system may use deep-learning style models such a neural network with multiple hidden layers.

[0022] Figure 2 is a flow diagram that illustrates the processing of a generate model component of the MPP system in some embodiments. The generate model component 200 is invoked to generate the model for a molecule for use in predictions. In block 201, the component invokes the collect structural feature information component to collect the structural features of the molecule. In block 202, the component invokes the collect variant

information component to collect variant information for each variant that is used to generate the model. In blocks 203-206, the component loops generating the feature vector for each variant. In block 203, the component selects the next variant. In decision block 204, if all variants have already selected, then the component continues at block 207, else the component continues at block 205. In block 205, the component invokes the generate feature vector component passing an indication of the variant. In block 206, the component assigns the value of the molecular property for the variant to the feature vector to form the training data of the feature vector and the assigned value. The component then loops to block 203 to select the next variant. In block 207, the component trains the prediction model using the training data and completes.

[0023] Figure 3 is a flow diagram that illustrates the processing of a collect structural feature information component of the MPP system in some embodiments. The collect structural feature information component 300 is invoked to collect the structural features for the molecule. In block 301, the component selects the next residue of the molecule. In decision block 302, if all the residues have already been selected, then the component completes, else the component continues at block 303. In block 303, the component selects the next structural feature of the residue. In decision block 304, if all the structural features for the selected residue have already been selected, then the component loops to block 301 to select the next residue, else the component continues at block 305. In block 305, the component accesses the value for the structural feature for the selected residue. In block 306, the component stores the value in the structural matrix 133 and then loops to block 303 to select the next structural feature.

[0024] Figure 4 is a flow diagram that illustrates the processing of a collect variant information component of the MPP system in some embodiments. The collect variant information component 400 is invoked to collect the molecular properties of the variants along with an indication of which residues were modified. In block 401, the component selects the next variant. In decision block 402, if all the variants have already been selected, then the component completes, else the component continues at block 403. In block 403, the component selects the next residue of the molecule. In decision block 404, if all the residues have already been selected, then the component loops to block 401 to

select the next variant, else the component continues at block 405. In decision block 405, if the residue is modified in the variant, then the component continues at block 406, else the component continues at block 407. In block 406, the component sets the entry for the selected variant and the selected residue in the residue matrix 132 to indicate that the residue has been modified in the selected variant. In block 407, the component selects the next molecular property. In decision block 408, if all the molecular properties have already been selected, then the component loops to block 403 to select the next residue in the molecule, else the component continues at block 409. In block 409, the component accesses the value for the molecular property. In block 410, the component stores the value in the molecular property matrix 131 and then loops to block 407 to select the next molecular property.

[0025] Figure 5 is a flow diagram that illustrates the processing of a generate feature vector component of the MPP system in some embodiments. A generate feature vector component 500 is invoked to generate a feature vector for a passed variant. In block 501, the component selects the next structural feature. In decision block 502, if all the structural features have already been selected, then the component completes, else the component continues at block 503. In block 503, the component selects the next modified residue for the passed variant. In decision block 504, if all the modified residues have already been selected, then the component continues at block 507, else the component continues at block 505. In block 505, the component accesses the value for the selected structural feature. In block 506, the component stores the value in the modified residue structural matrix 134 for the variant and then loops to block 503 to select the next modified residue. In block 507, the component invokes a generate feature component to generate the features for the variant from the value for the selected structural feature stored in the modified residue structural matrix 134 and then loops to block 501 to select the next structural feature.

[0026] Figure 6 is a flow diagram that illustrates the processing of a generate feature component of the MPP system in some embodiments. The generate feature component 600 is passed an indication of a variant and a structural feature and generates a feature for the feature vector for the variant for each statistic based on the structural feature. In block

601, the component selects the next statistic. In decision block 602, if all the statistics have already been selected, then the component completes, else the component continues at block 603. In block 603, the component generates the selected statistic for the structural feature of the variant based on values in the modified residue structural matrix 134 for the passed variant. In block 604, the component stores the statistic in the structural feature summary matrix 135 and then loops to block 601 to select the next statistic.

[0027] Figure 7 is a flow diagram that illustrates the processing of a predict molecular property component of the MPP system in some embodiments. The predict property component 700 is invoked to predict a molecular property of a new variant. The new variant is indicated by the residues that have been modified. In block 701, the component invokes a generate feature vector component to generate a feature vector for the new variant based on the modified residues as indicated by the residue array 151. In block 702, the component applies the prediction model to the feature vector to generate a value for a molecular property for the new variant. In block 703, the component outputs of value of the molecular property and then completes.

[0028] The following paragraphs describe various embodiments of aspects of the MPP system. An implementation of the MPP system may employ any combination of the embodiments. The processing described below may be performed by a computing device with a processor that executes computer-executable instructions stored on a computer-readable storage medium that implements the MPP system.

[0029] In some embodiments, a method performed by a computing system for generating a model for predicting a molecular property of a variant of a molecule is provided. The method accesses values for structural features of residues of the molecule. For each of a plurality of variants of the molecule, the method accesses variant information indicating which residues in a sequence of residues of the molecule were modified to form the variant and a value for the molecular property of the variant. For each of the plurality of variants of the molecule, the method also, for each structural feature, aggregates the values for the structural features of the residues of the molecule that were modified to form the variant to form a feature vector for the variant. For each of the plurality of variants of

the molecule, the method assigns the value for the molecular property of the variant to the feature vector wherein the feature vector and the assigned value form training data. The method then generates the model for predicting a value for the molecular property using the training data for the plurality of variants. In some embodiments, the method further predicts a value for the molecular property of a new variant by accessing new variant information indicating which residues in the sequence of residues of the molecule were modified to form the variant; for each structural feature, aggregates the values for the structural feature of the residues of the molecule that were modified to form the new variant to form a new feature vector for the new variant; and applies the model to the new feature vector to predict the value for the molecular property of the new variant. In some embodiments, the model is generated using a linear regression technique using the training data as input. In some embodiments, the model is generated by learning a neural network using the training data as input. In some embodiments, the generating of the model includes reducing dimensions of the training data. In some embodiments, the molecule is a protein. In some embodiments, a variant is formed by replacing an amino acid of the molecule with a different amino acid. In some embodiments, the molecular property is selected from a group consisting of antibody melting temperature, percentage of high molecular weight, chemical unfolding behavior, solubility, viscosity, and aggregation behavior. In some embodiments, the structural features are selected from a group consisting of participating of a residue in charge patches or hydrophobic patches, group diversity of neighboring residues, solvent accessibility of a residue, nearness to region/chain interfaces, secondary structural environment, sizes of residue in the molecule and the variant, and acidity of a residue in the molecule and the variant. In some embodiments, the aggregating of the values for a structural feature generates statistics selected from a group consisting of sum, mean, standard deviation, skew, kurtosis, minimum, maximum, product, sum of log of absolute values, and mean of log of absolute values.

[0030] In some embodiments, a computing system for predicting a value for a molecular property of a new variant of a molecule is provided. The computing system includes one or more computer-readable storage medium storing computer-executable instructions and one or more processors for executing the computer-executable

instructions stored in the one or more computer-readable mediums. The computer-executable instructions control the computing system to access new variant information indicating which residues in a sequence of residues of the molecule were modified to form the new variant. For each of a plurality of structural features of residues of the molecule, the computer-executable instructions control the computing system to aggregate the values for the structural feature of the residues of the molecule that were modified to form the new variant to form a new feature vector for the new variant. The computer-executable instructions further control the computing system apply a model to the new feature vector to predict the value for the molecular property of the new variant. The model is generated using training data comprising feature vectors derived from value of structural features of variants of the molecule and the values of the molecular property of those variants. In some embodiments, the computer-executable instructions further control the computing system to access values for structural features of residues of the molecule. For each of a plurality of variants of the molecule, the computer-executable instructions control the computing system to access variant information indicating which residues in a sequence of residues of the molecule were modified to form the variant and a value for the molecular property of the variant; for each structural feature, aggregate the values for the structural features of the residues of the molecule that were modified to form the variant to form the feature vector for the variant; and assign the value for the molecular property of the variant to the feature vector wherein the feature vector and the assigned value form the training data. The computer-executable instructions control the computing system to generate the model for predicting a value for the molecular property using the training data for the plurality of variants. In some embodiments, the model is generated using a linear regression technique using the training data as input. In some embodiments, the model is generated by learning a neural network using the training data as input. In some embodiments, the computer-executable instructions further control the computing system to reduce dimensions of the training data. In some embodiments, the molecule is a protein. In some embodiments, a variant is formed by replacing an amino acid of the molecule with a different amino acid. In some embodiments, the molecular property is selected from a group consisting of antibody melting temperature, percentage of high molecular weight, chemical unfolding behavior, solubility, viscosity, and aggregation

behavior. In some embodiments, the structural features are selected from a group consisting of participating of a residue in charge patches or hydrophobic patches, group diversity of neighboring residues, solvent accessibility of a residue, nearness to region/chain interfaces, secondary structural environment, sizes of residue in the molecule and the variant, and acidity of a residue in the molecule and the variant. In some embodiments, the computer-executable instructions control the computing system to aggregate the values for a structural feature further generate statistics selected from a group consisting of sum, mean, standard deviation, skew, kurtosis, minimum, maximum, product, sum of log of absolute values, and mean of log of absolute values.

[0031] Although the subject matter has been described in language specific to structural features and/or acts, it is to be understood that the subject matter defined in the appended claims is not necessarily limited to the specific features or acts described above. Rather, the specific features and acts described above are disclosed as example forms of implementing the claims. Accordingly, the invention is not limited except as by the appended claims.

CLAIMS

I/We claim:

1. A method performed by a computing system for generating a model for predicting a molecular property of a variant of a molecule, the method comprising:

accessing values for structural features of residues of the molecule;

for each of a plurality of variants of the molecule,

accessing variant information indicating which residues in a sequence of residues of the molecule were modified to form the variant and a value for the molecular property of the variant;

for each structural feature, aggregating the values for the structural features of the residues of the molecule that were modified to form the variant to form a feature vector for the variant; and

assigning the value for the molecular property of the variant to the feature vector wherein the feature vector and the assigned value form training data; and

generating the model for predicting a value for the molecular property using the training data for the plurality of variants.

2. The method of claim 1 further comprising predicting a value for the molecular property of a new variant by:

accessing new variant information indicating which residues in the sequence of residues of the molecule were modified to form the variant;

for each structural feature, aggregating the values for the structural feature of the residues of the molecule that were modified to form the new variant to form a new feature vector for the new variant; and

applying the model to the new feature vector to predict the value for the molecular property of the new variant.

3. The method of claim 1 wherein the model is generated using a linear regression technique using the training data as input.

4. The method of claim 1 wherein the model is generated by learning a neural network using the training data as input.

5. The method of claim 1 wherein the generating of the model includes reducing dimensions of the training data.

6. The method of claim 1 wherein the molecule is a protein.

7. The method of claim 1 wherein a variant is formed by replacing an amino acid of the molecule with a different amino acid.

8. The method of claim 1 wherein the molecular property is selected from a group consisting of antibody melting temperature, percentage of high molecular weight, chemical unfolding behavior, solubility, viscosity, and aggregation behavior.

9. The method of claim 1 wherein the structural features are selected from a group consisting of participating of a residue in charge patches or hydrophobic patches, group diversity of neighboring residues, solvent accessibility of a residue, nearness to region/chain interfaces, secondary structural environment, sizes of residue in the molecule and the variant, and acidity of a residue in the molecule and the variant.

10. The method of claim 1 wherein the aggregating of the values for a structural feature generates statistics selected from a group consisting of sum, mean, standard deviation, skew, kurtosis, minimum, maximum, product, sum of log of absolute values, and mean of log of absolute values.

11. A computing system for predicting a value for a molecular property of a new variant of a molecule, the computing system comprising:

one or more computer-readable storage medium storing computer-executable instructions for controlling the computing system to:

access new variant information indicating which residues in a sequence of residues of the molecule were modified to form the new variant;

for each of a plurality of structural features of residues of the molecule, aggregate the values for the structural feature of the residues of the molecule that were modified to form the new variant to form a new feature vector for the new variant; and

applying a model to the new feature vector to predict the value for the molecular property of the new variant, the model being generated using training data comprising feature vectors derived from value of structural features of variants of the molecule and the values of the molecular property of those variants; and

one or more processors for executing the computer-executable instructions stored in the one or more computer-readable mediums.

12. The computing system of claim 11 wherein the computer-executable instructions further control the computing system to:

access values for structural features of residues of the molecule;

for each of a plurality of variants of the molecule,

access variant information indicating which residues in a sequence of residues of the molecule were modified to form the variant and a value for the molecular property of the variant;

for each structural feature, aggregate the values for the structural features of the residues of the molecule that were modified to form the variant to form the feature vector for the variant; and

assign the value for the molecular property of the variant to the feature vector wherein the feature vector and the assigned value form the training data; and

generate the model for predicting a value for the molecular property using the training data for the plurality of variants.

13. The computing system of claim 12 wherein the model is generated using a linear regression technique using the training data as input.

14. The computing system of claim 12 wherein the model is generated by learning a neural network using the training data as input.

15. The computing system of claim 12 wherein the computer-executable instructions further control the computing system to reduce dimensions of the training data.

16. The computing system of claim 11 wherein the molecule is a protein.

17. The computing system of claim 11 wherein a variant is formed by replacing an amino acid of the molecule with a different amino acid.

18. The computing system of claim 11 wherein the molecular property is selected from a group consisting of antibody melting temperature, percentage of high molecular weight, chemical unfolding behavior, solubility, viscosity, and aggregation behavior.

19. The computing system of claim 11 wherein the structural features are selected from a group consisting of participating of a residue in charge patches or hydrophobic patches, group diversity of neighboring residues, solvent accessibility of a residue, nearness to region/chain interfaces, secondary structural environment, sizes of residue in the molecule and the variant, and acidity of a residue in the molecule and the variant.

20. The computing system of claim 11 wherein the computer-executable instructions control the computing system to aggregate the values for a structural feature

further generate statistics selected from a group consisting of sum, mean, standard deviation, skew, kurtosis, minimum, maximum, product, sum of log of absolute values, and mean of log of absolute values.

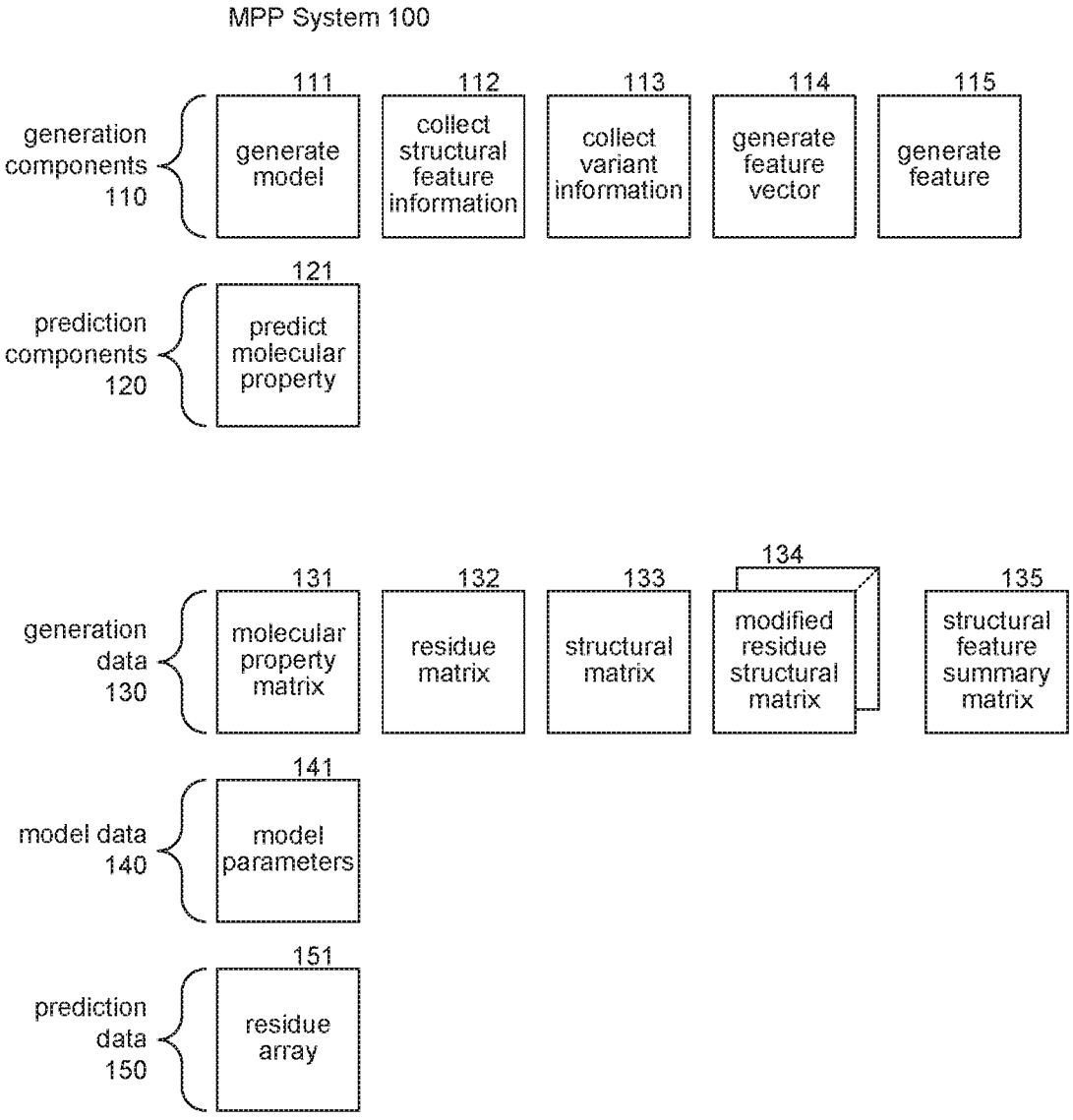
MPP System 100
generation components 110
111
generate model
112
collect structural feature information
113
collect variant information
114
generate feature vector
115
generate feature
prediction components 120
121
predict molecular property
generation data 130
131
molecular property matrix
132
residue matrix
133
structural matrix
134
modified residue structural matrix
135
structural feature summary matrix
model data 140
141
model parameters
prediction data 150
151
residue array

FIG. 1

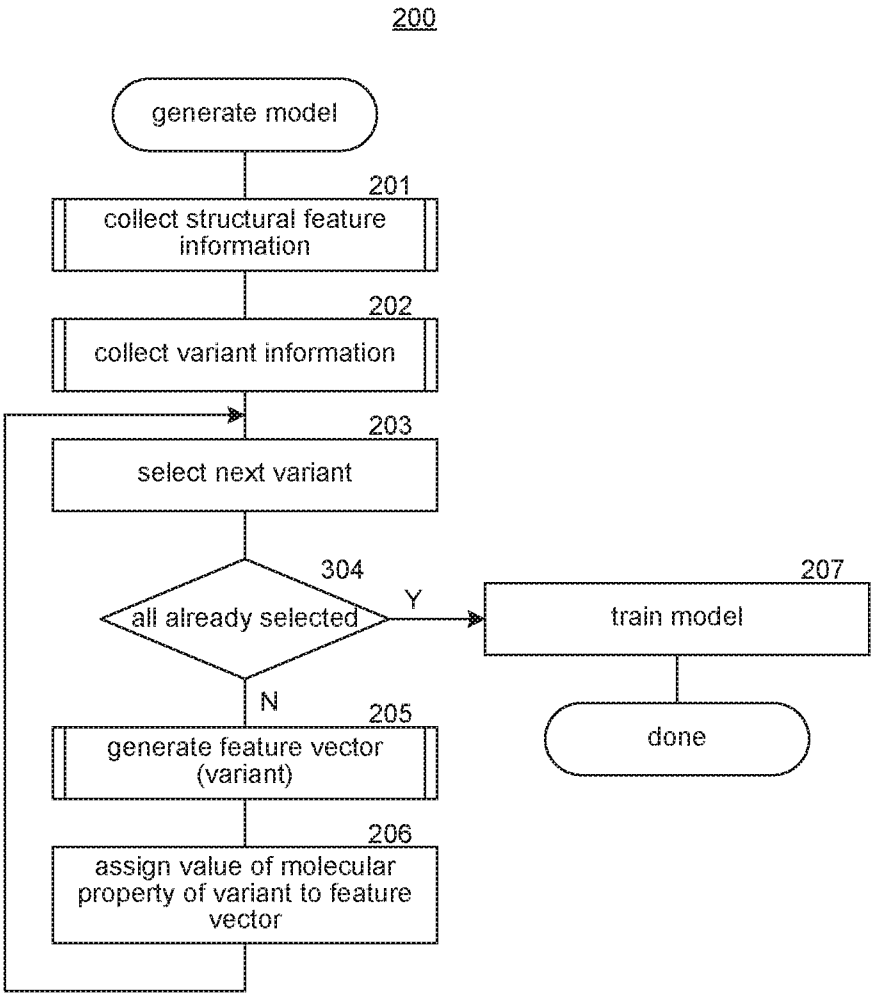
200
generate model
201
collect structural feature information
202
collect variant information
203
select next variant
304
all already selected
Y
207
train model
done
N
205
generate feature vector (variant)
206
assign value of molecular property of variant to feature vector

FIG. 2

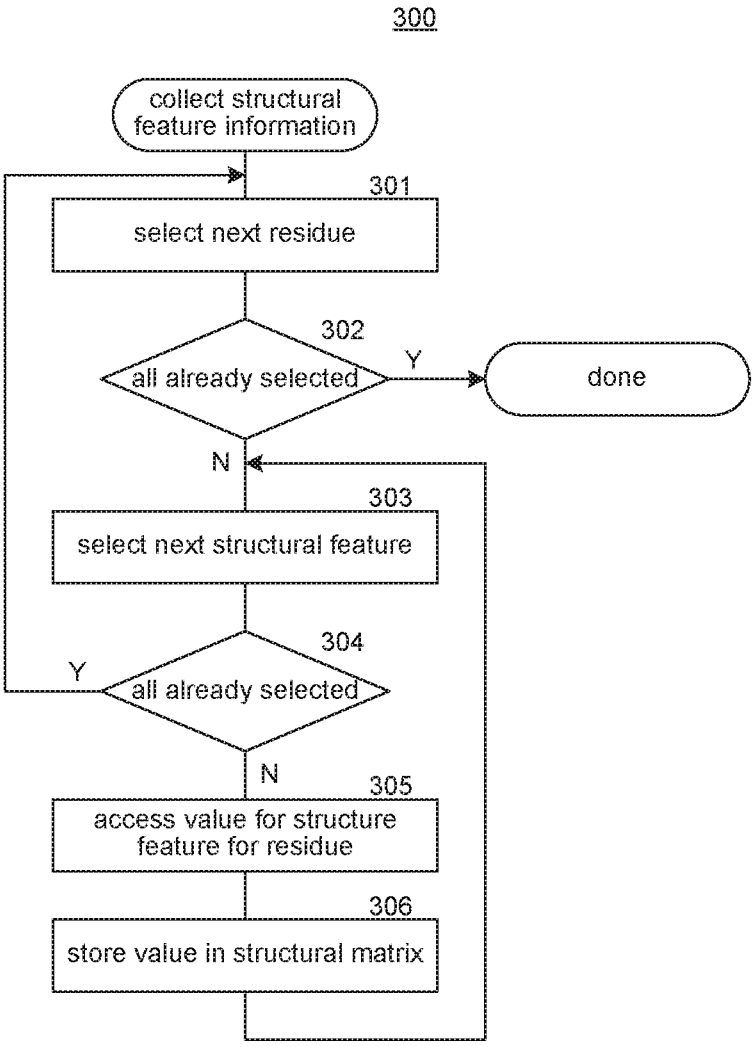
300
collect structural feature information
301
select next residue
302
all already selected
Y
done
N
303
select next structural feature
304
all already selected
Y
N
305
access value for structure feature for residue
306
store value in structural matrix

FIG. 3

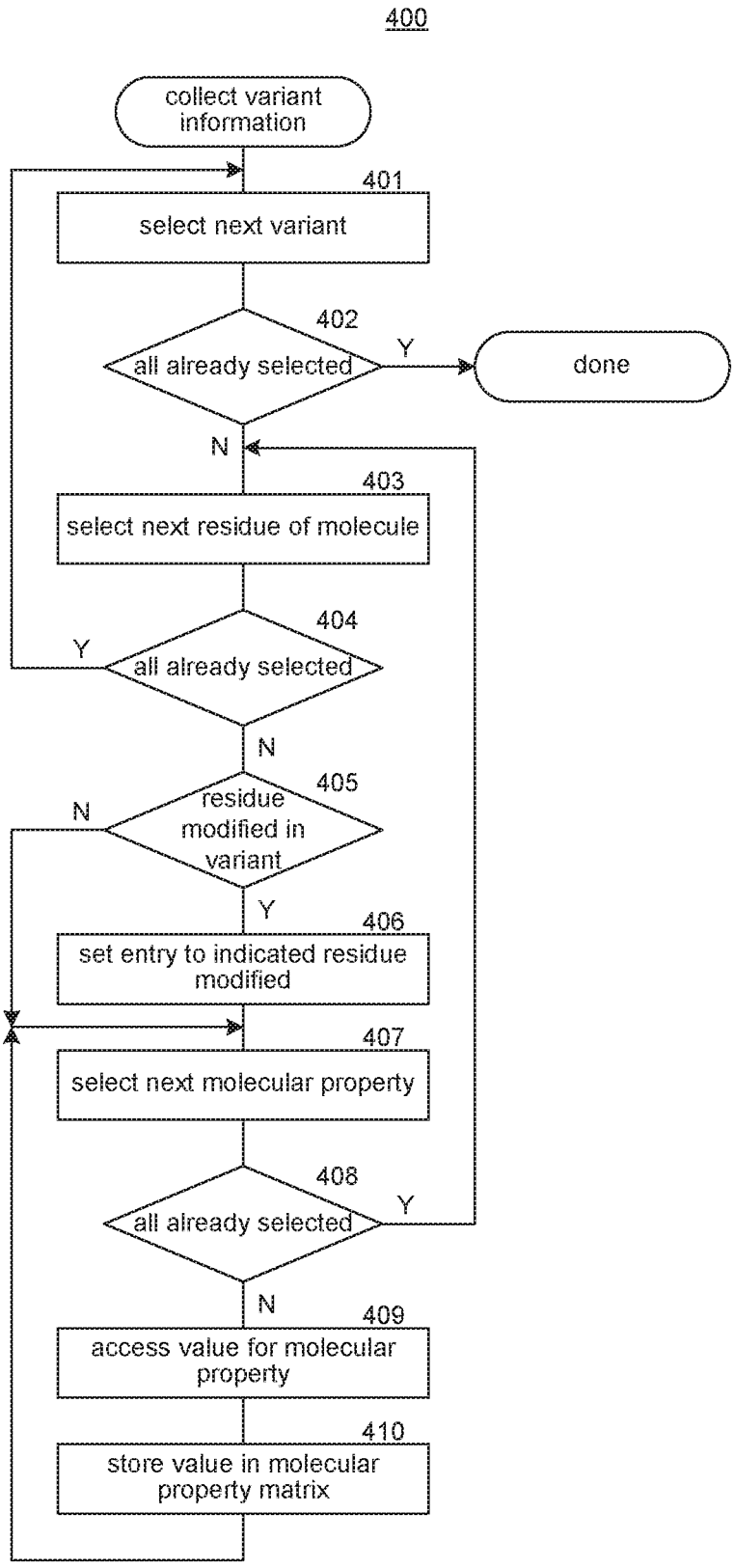
400
collect variant information
401
select next variant
402
all already selected
Y
done
N
403
select next residue of molecule
404
all already selected
Y
N
405
residue modified in variant
N
Y
406
set entry to indicated residue modified
407
select next molecular property
408
all already selected
Y
N
409
access value for molecular property
410
store value in molecular property matrix

FIG. 4

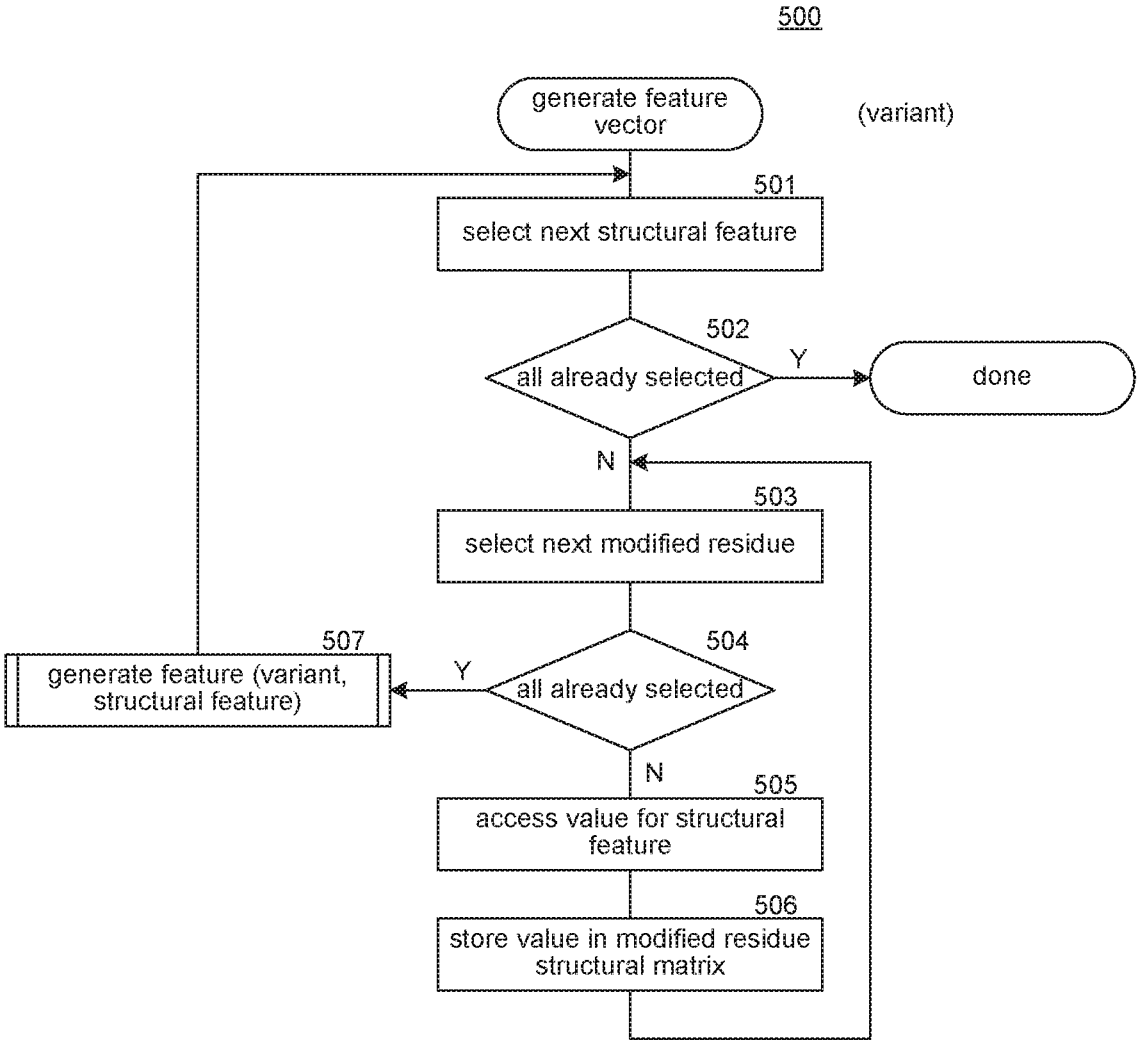
500
generate feature vector
(variant)
501
select next structural feature
502
all already selected
Y
done
N
503
select next modified residue
504
all already selected
Y
507
generate feature (variant, structural feature)
N
505
access value for structural feature
506
store value in modified residue structural matrix

FIG. 5

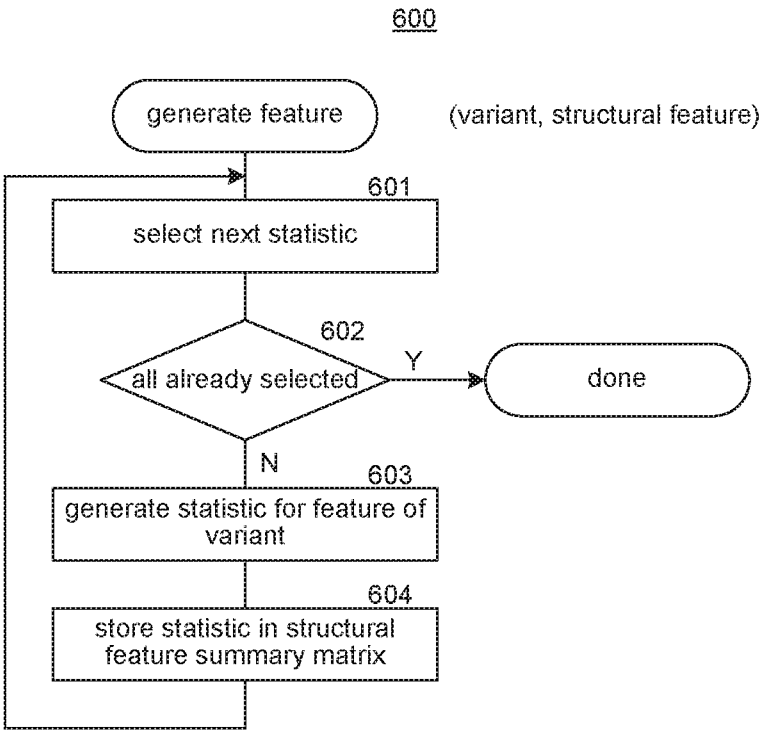
600
generate feature
(variant, structural feature)
601
select next statistic
602
all already selected
Y
done
N
603
generate statistic for feature of variant
604
store statistic in structural feature summary matrix

FIG. 6

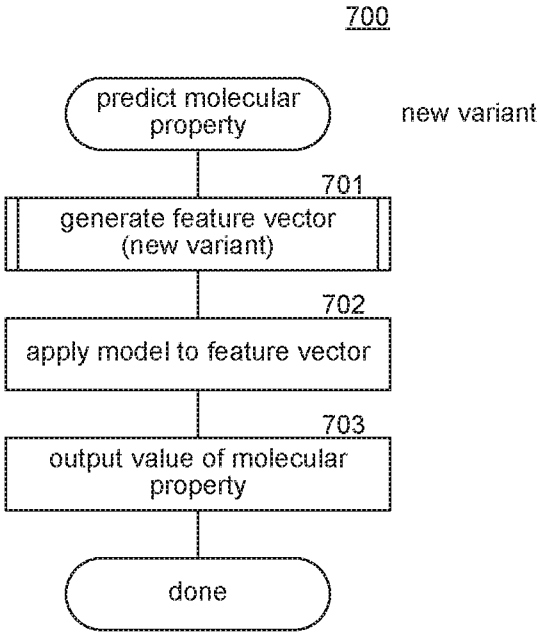
700
predict molecular property
new variant
701
generate feature vector (new variant)
702
apply model to feature vector
703
output value of molecular property
done

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2018/036777

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/18 G06F19/24 C40B30/02 C40B30/10
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F C40B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	OLGA OBREZANOVA ET AL: "Aggregation risk prediction for antibodies and its application to biotherapeutic development", MABS, vol. 7, no. 2, 4 March 2015 (2015-03-04), pages 352-363, XP055268078, US ISSN: 1942-0862, DOI: 10.1080/19420862.2015.1007828 p. 353, left-hand column, 3rd paragraph p. 353, right-hand column 1st paragraph p. 353, right-hand column, 2nd paragraph p. 355, left-hand column, section "Results", sub-section "Physico-chemical sequence descriptors": p. 353, section "Results", sub-section "Antibody set design" p. 354 to 355, section "Results", -/--	1-20

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
7 September 2018	18/09/2018

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Heidrich, Alexander

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	sub-section "Expression and analytical characterization" p. 355, section "Results", sub-section "Model construction and validation" p. 357 to 358, section "Results", sub-section "Application of aggregation tool for selection of antibody variants" p. 360, left-hand column, section "Materials and Methods", sub-section "Physico-chemical sequence descriptios" p. 362, section "Access to Software" sub-sections "Performance of predictive models" (p. 355 to 356) and "Model validation on external data" (p. 356 to 357) -----	
X	YANG YANG ET AL: "PON-Sol: prediction of effects of amino acid substitutions on protein solubility", BIOINFORMATICS., vol. 32, no. 13, 19 February 2016 (2016-02-19), pages 2032-2034, XP055496382, GB ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btw066 the whole document -----	1-20
X	LIN H ET AL: "Prediction of thermophilic proteins using feature selection technique", JOURNAL OF MICROBIOLOGICAL METHODS, ELSEVIER, AMSTERDAM, NL, vol. 84, no. 1, 31 October 2010 (2010-10-31), pages 67-70, XP027563983, ISSN: 0167-7012 [retrieved on 2010-10-31] the whole document -----	1-20
X	HAO LIN ET AL: "AcalPred: A Sequence-Based Tool for Discriminating between Acidic and Alkaline Enzymes", PLOS ONE, vol. 8, no. 10, 9 October 2013 (2013-10-09), page e75726, XP055496584, DOI: 10.1371/journal.pone.0075726 the whole document -----	1-20
A	US 2005/278124 A1 (DUFFY NIGEL P [US] ET AL) 15 December 2005 (2005-12-15) the whole document -----	1-20
	-/--	

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2018/036777

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	GERARD JP VAN WESTEN ET AL: "Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets", JOURNAL OF CHEMINFORMATICS, BIOMED CENTRAL LTD, LONDON, UK, vol. 5, no. 1, 23 September 2013 (2013-09-23), page 41, XP021162263, ISSN: 1758-2946, DOI: 10.1186/1758-2946-5-41 the whole document -----	1-20
A	L. MONTANUCCI ET AL: "Predicting protein thermostability changes from sequence upon multiple mutations", BIOINFORMATICS., vol. 24, no. 13, 27 June 2008 (2008-06-27) , pages i190-i195, XP055496848, GB ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btn166 the whole document, in particula section "Introduction" -----	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2018/036777

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2005278124 A1	15-12-2005	NONE	