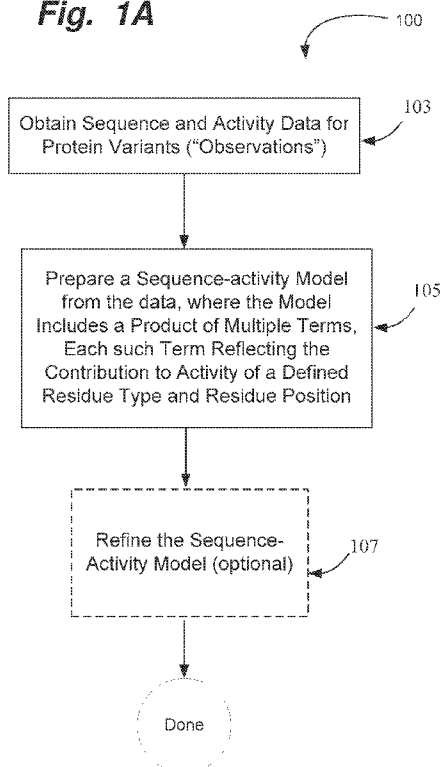




- (51) **International Patent Classification:**
G01N 33/50 (2006.01)
- (21) **International Application Number:**
PCT/US2014/013668
- (22) **International Filing Date:**
29 January 2014 (29.01.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
61/759,276 31 January 2013 (31.01.2013) US
61/799,377 15 March 2013 (15.03.2013) US
- (71) **Applicant (for all designated States except US):** **CODEX-IS, INC.** [US/US]; 200 Penobscot Drive, Redwood City, California 94063 (US).
- (72) **Inventors; and**
- (71) **Applicants (for US only):** **COPE, Gregory Allan** [US/US]; 200 Penobscot Drive, Redwood City, California 94603 (US). **AGARD, Nicholas John** [US/US]; 200 Penobscot Drive, Redwood City, California 94063 (US).
- (74) **Agents:** **WEAVER, Jeffrey, K.** et al.; Weaver Austin Villeneuve & Sampson LLP, P.O. Box 70250, Oakland, California 94612-0250 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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.

[Continued on next page]

(54) **Title:** METHODS, SYSTEMS, AND SOFTWARE FOR IDENTIFYING BIO-MOLECULES USING MODELS OF MULTIPLICATIVE FORM

Fig. 1A

(57) **Abstract:** 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 comprising multiplicative terms and using the models to guide directed evolution. In some embodiments, the sequence-activity models include one or more interaction terms, each of which including an interaction coefficient representing the contribution to activity of two or more defined residues. In some embodiments, the models describe relation between protein or nucleic acid sequences and protein activities. In some embodiments, the present invention also provides methods for preparing sequence-activity models, including but not limited to stepwise addition or subtraction techniques, Bayesian regression, ensemble regression and other methods. The present invention further provides digital systems and software for performing the methods provided herein.



(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, SZ, TZ, 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))*