**Predicting serious rare adverse reactions of novel chemicals**

Aleksandar Poleksic and Lei Xie

**Supplementary Material**

**CS algorithm for drug-ADR association prediction**

**Goal:** Predict missing drug-ADR associations.

**Input:** An incomplete binary matrix of known (observed) associations , a matrix containing frequencies of drug-ADR associations, an impute matrix , a matrix of pairwise ADR similarity scores, a matrix of pairwise drug similarity scores.

**Output:** The matrix of probabilities of individual drug-ADR associations.

**Implementation details:**

Denote by and the number of ADRs and drugs respectively and let be a matrix of drug-ADR associations ( if drug is known to cause ADR and otherwise). The binary matrix is derived from the SIDER database and is incomplete in the sense that a zero entry does not necessarily mean that is not the adverse reaction of , but it simply indicates the lack of evidence of any such association.

Our algorithm attempts to complete by reclassifying some of its entries. The key assumption is that the true matrix of drug-ADR associations is of low rank and, therefore, it can be written as the product of two matrices and of dimensions and , respectively, where .

Following Steck (2010) and Lim *et al.* (2016), we compute the matrices and by minimizing the loss function:

+ (1)

The meaning of each term of (1) is provided in the Methods section of the main text. We note that the partial derivatives of (1) are given by:

where represents the Hadamard product. The matrices and satisfying and are found in an iterative fashion using the gradient descent method Ada-Grad (Duchi et al., 2011).

To address the so-called “cold-start” problem i.e., to predict drug-ADR associations for new drugs (those without any observed ADRs) we use the so-called “weighted-profile” method (Yamanishi *et al.*, 2008; Yao *et al.* 2014). More precisely, the row of , containing the latent preferences of the drug , is computed as the weighted sum of latent preferences for the drug (calculated during the iterative minimization procedure, described above) and the latent preferences of drugs most similar to the drug . More specifically, the row of the matrix is set to , where is the row of (representing the latent preferences of the drug ), is the weight parameter and represents the Tanimoto similarity of drugs and . The normalization factor is set to .

 In case of completely new drugs, the above scores are further multiplied by the prior and posterior ADR probabilities computed from the training set in a manner similar to that used in the ML algorithm (see the definitions of and given in the main text). We use the same “weighted-profile” method to predict drugs for “knew” ADRs.

 The parameters of the loss function (1) are trained on a selected matrix of drug-ADR associations derived from the post-marketing data. This data is available as part of the SIDER database. The number of AdaGrad iterations and the matrix rank are each set to 100. The parameter is set to 5. In our experience, the choice of the algorithm’s parameters do not significantly affect the algorithm’s performance, as long as the parameters were selected within a reasonable range (e.g. between 0.01 and 1, number of iterations above 50 and between 3 and 10). In fact, due to the complexity of the training procedure, some methods for similar or unrelated problems pick these parameters somewhat arbitrary (Yao *et al*., 2014).

****

**Figure S1.** **Predicting ADRs responsible for drug market withdrawals.** The x-axis represents the percentage (0%, 10%, 25%, 50%) of already known ADRs for the drug. The y-axis represents the AUC values. The mean AUC values shown in the figures are obtained over multiple runs on “control” drugs to achieve robust statistics (STDERR too small to show).



**Figure S2.** **Value added by compressed sensing (CS) in MCC benchmark.** The data table beneath the graph shows the mean MCC scores obtained in 5 rounds of 10-fold CV benchmark. All STDERR values are smaller than .



**Figure S3.** **Value added by compressed sensing (CS) in the “cold-start” setting.** The mean MCC scores obtained in 5 rounds of the 10-fold cross-validation on the set of “new” drugs, those with all ADRs hidden (masked out). All STDERR values are smaller than .



**Figure S4.** **MCC test for rare ADRs.** The mean MCC values obtained in 5 rounds of 10-fold CV test on the set of drugs with no known rare ADRs. All STDERR values are smaller than .

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **CS** |  | **ML** |  | **CCA** |
| **CutOff** | **AUC** | **STDERR** | **ENRCH** |  | **AUC** | **STDERR** | **ENRCH** |  | **AUC** | **STDERR** | **ENRCH** |
| **<12 drugs per ADR** |
| **0.1** | 0.769 | 0.002 | 1.5 |  | 0.505 | 0.005 | 1.0 |  | 0.525 | 0.005 | 1.1 |
| **0.2** | 0.790 | 0.004 | 1.6 |  | 0.556 | 0.007 | 1.1 |  | 0.563 | 0.006 | 1.1 |
| **0.3** | 0.805 | 0.003 | 1.6 |  | 0.576 | 0.004 | 1.2 |  | 0.574 | 0.006 | 1.1 |
| **0.4** | 0.823 | 0.002 | 1.6 |  | 0.590 | 0.004 | 1.2 |  | 0.587 | 0.004 | 1.2 |
| **0.5** | 0.833 | 0.003 | 1.7 |  | 0.608 | 0.005 | 1.2 |  | 0.605 | 0.006 | 1.2 |
| **0.6** | 0.831 | 0.002 | 1.7 |  | 0.613 | 0.003 | 1.2 |  | 0.615 | 0.003 | 1.2 |
| **0.7** | 0.841 | 0.004 | 1.7 |  | 0.626 | 0.006 | 1.3 |  | 0.625 | 0.006 | 1.3 |
| **0.8** | 0.846 | 0.003 | 1.7 |  | 0.644 | 0.004 | 1.3 |  | 0.634 | 0.005 | 1.3 |
| **0.9** | 0.853 | 0.002 | 1.7 |  | 0.660 | 0.009 | 1.3 |  | 0.641 | 0.010 | 1.3 |
| **1** | 0.856 | 0.006 | 1.7 |  | 0.666 | 0.005 | 1.3 |  | 0.639 | 0.004 | 1.3 |
| **<25 drugs per ADR** |
| **0.1** | 0.768 | 0.001 | 1.5 |  | 0.608 | 0.002 | 1.2 |  | 0.621 | 0.002 | 1.2 |
| **0.2** | 0.803 | 0.005 | 1.6 |  | 0.645 | 0.006 | 1.3 |  | 0.654 | 0.006 | 1.3 |
| **0.3** | 0.820 | 0.003 | 1.6 |  | 0.671 | 0.004 | 1.3 |  | 0.673 | 0.005 | 1.3 |
| **0.4** | 0.834 | 0.003 | 1.7 |  | 0.678 | 0.006 | 1.4 |  | 0.682 | 0.006 | 1.4 |
| **0.5** | 0.844 | 0.002 | 1.7 |  | 0.701 | 0.004 | 1.4 |  | 0.700 | 0.005 | 1.4 |
| **0.6** | 0.849 | 0.006 | 1.7 |  | 0.695 | 0.005 | 1.4 |  | 0.696 | 0.005 | 1.4 |
| **0.7** | 0.861 | 0.002 | 1.7 |  | 0.712 | 0.002 | 1.4 |  | 0.709 | 0.004 | 1.4 |
| **0.8** | 0.859 | 0.001 | 1.7 |  | 0.720 | 0.003 | 1.4 |  | 0.711 | 0.004 | 1.4 |
| **0.9** | 0.871 | 0.002 | 1.7 |  | 0.751 | 0.004 | 1.5 |  | 0.731 | 0.005 | 1.5 |
| **1** | 0.873 | 0.003 | 1.7 |  | 0.757 | 0.002 | 1.5 |  | 0.732 | 0.002 | 1.5 |
| **<50 drugs per ADR** |
| **0.1** | 0.738 | 0.002 | 1.5 |  | 0.656 | 0.002 | 1.3 |  | 0.667 | 0.001 | 1.3 |
| **0.2** | 0.811 | 0.003 | 1.6 |  | 0.728 | 0.004 | 1.5 |  | 0.736 | 0.004 | 1.5 |
| **0.3** | 0.835 | 0.003 | 1.7 |  | 0.745 | 0.005 | 1.5 |  | 0.749 | 0.005 | 1.5 |
| **0.4** | 0.840 | 0.002 | 1.7 |  | 0.745 | 0.004 | 1.5 |  | 0.748 | 0.004 | 1.5 |
| **0.5** | 0.856 | 0.003 | 1.7 |  | 0.752 | 0.004 | 1.5 |  | 0.755 | 0.003 | 1.5 |
| **0.6** | 0.864 | 0.001 | 1.7 |  | 0.756 | 0.004 | 1.5 |  | 0.756 | 0.006 | 1.5 |
| **0.7** | 0.867 | 0.003 | 1.7 |  | 0.766 | 0.004 | 1.5 |  | 0.763 | 0.004 | 1.5 |
| **0.8** | 0.877 | 0.003 | 1.8 |  | 0.780 | 0.003 | 1.6 |  | 0.771 | 0.003 | 1.5 |
| **0.9** | 0.881 | 0.004 | 1.8 |  | 0.795 | 0.005 | 1.6 |  | 0.779 | 0.004 | 1.6 |
| **1** | 0.883 | 0.003 | 1.8 |  | 0.812 | 0.003 | 1.6 |  | 0.787 | 0.003 | 1.6 |
| **<100 drugs per ADR** |
| **0.1** | 0.779 | 0.001 | 1.6 |  | 0.669 | 0.003 | 1.3 |  | 0.678 | 0.002 | 1.4 |
| **0.2** | 0.782 | 0.003 | 1.6 |  | 0.785 | 0.003 | 1.6 |  | 0.789 | 0.004 | 1.6 |
| **0.3** | 0.835 | 0.002 | 1.7 |  | 0.807 | 0.002 | 1.6 |  | 0.810 | 0.002 | 1.6 |
| **0.4** | 0.860 | 0.003 | 1.7 |  | 0.803 | 0.003 | 1.6 |  | 0.806 | 0.003 | 1.6 |
| **0.5** | 0.868 | 0.002 | 1.7 |  | 0.808 | 0.003 | 1.6 |  | 0.806 | 0.004 | 1.6 |
| **0.6** | 0.876 | 0.005 | 1.8 |  | 0.811 | 0.003 | 1.6 |  | 0.813 | 0.004 | 1.6 |
| **0.7** | 0.885 | 0.001 | 1.8 |  | 0.820 | 0.001 | 1.6 |  | 0.814 | 0.002 | 1.6 |
| **0.8** | 0.897 | 0.003 | 1.8 |  | 0.843 | 0.002 | 1.7 |  | 0.832 | 0.003 | 1.7 |
| **0.9** | 0.893 | 0.002 | 1.8 |  | 0.841 | 0.003 | 1.7 |  | 0.826 | 0.003 | 1.7 |
| **1** | 0.901 | 0.002 | 1.8 |  | 0.854 | 0.004 | 1.7 |  | 0.833 | 0.004 | 1.7 |

**Table S1.** **The results of LOOCV for chemicals of novel 3D structure using AUC as the accuracy measure.** The AUC scores shown are the mean values obtained from 5 rounds of the leave-one-out-cross-validation test on the sets of 100 randomly chosen drugs. For each drug tested, we removed from the training set all drugs form the same structural class i.e., those that have the above cutoff Tanimoto similarity to the test drug. ENRCH columns gives the fold enrichment over the random classifier.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **CS** |  | **ML** |  | **CCA** |
| **CutOff** | **AUPR** | **STDERR** | **ENRCH** |  | **AUPR** | **STDERR** | **ENRCH** |  | **AUPR** | **STDERR** | **ENRCH** |
| **<12 drugs per ADR** |
| **0.1** | 0.228 | 0.011 | 2.9 |  | 0.101 | 0.008 | 1.3 |  | 0.116 | 0.010 | 1.5 |
| **0.2** | 0.120 | 0.005 | 5.1 |  | 0.037 | 0.003 | 1.6 |  | 0.043 | 0.002 | 1.8 |
| **0.3** | 0.097 | 0.006 | 8.2 |  | 0.020 | 0.001 | 1.7 |  | 0.019 | 0.001 | 1.6 |
| **0.4** | 0.082 | 0.004 | 11.9 |  | 0.015 | 0.003 | 2.2 |  | 0.011 | 0.001 | 1.5 |
| **0.5** | 0.082 | 0.005 | 14.4 |  | 0.014 | 0.002 | 2.4 |  | 0.009 | 0.000 | 1.6 |
| **0.6** | 0.088 | 0.004 | 15.9 |  | 0.016 | 0.002 | 2.9 |  | 0.009 | 0.000 | 1.7 |
| **0.7** | 0.091 | 0.007 | 16.6 |  | 0.016 | 0.002 | 3.0 |  | 0.010 | 0.000 | 1.8 |
| **0.8** | 0.106 | 0.006 | 19.8 |  | 0.030 | 0.001 | 5.6 |  | 0.011 | 0.000 | 2.0 |
| **0.9** | 0.134 | 0.007 | 26.8 |  | 0.052 | 0.006 | 10.5 |  | 0.012 | 0.001 | 2.3 |
| **1** | 0.120 | 0.008 | 24.4 |  | 0.055 | 0.004 | 11.1 |  | 0.011 | 0.001 | 2.3 |
| **<25 drugs per ADR** |
| **0.1** | 0.231 | 0.007 | 2.8 |  | 0.166 | 0.006 | 2.0 |  | 0.189 | 0.006 | 2.3 |
| **0.2** | 0.139 | 0.004 | 4.6 |  | 0.069 | 0.004 | 2.3 |  | 0.086 | 0.005 | 2.8 |
| **0.3** | 0.104 | 0.004 | 7.1 |  | 0.035 | 0.002 | 2.4 |  | 0.041 | 0.001 | 2.8 |
| **0.4** | 0.099 | 0.003 | 10.0 |  | 0.024 | 0.001 | 2.4 |  | 0.027 | 0.001 | 2.7 |
| **0.5** | 0.082 | 0.003 | 13.2 |  | 0.018 | 0.002 | 3.0 |  | 0.017 | 0.001 | 2.8 |
| **0.6** | 0.086 | 0.001 | 14.1 |  | 0.019 | 0.003 | 3.2 |  | 0.015 | 0.001 | 2.5 |
| **0.7** | 0.104 | 0.004 | 18.6 |  | 0.025 | 0.004 | 4.5 |  | 0.016 | 0.001 | 2.8 |
| **0.8** | 0.108 | 0.002 | 19.4 |  | 0.042 | 0.001 | 7.5 |  | 0.017 | 0.001 | 3.0 |
| **0.9** | 0.134 | 0.005 | 25.1 |  | 0.061 | 0.005 | 11.5 |  | 0.019 | 0.002 | 3.6 |
| **1** | 0.149 | 0.003 | 26.0 |  | 0.080 | 0.001 | 13.8 |  | 0.021 | 0.001 | 3.7 |
| **<50 drugs per ADR** |
| **0.1** | 0.237 | 0.007 | 2.7 |  | 0.230 | 0.007 | 2.6 |  | 0.230 | 0.007 | 2.6 |
| **0.2** | 0.175 | 0.007 | 4.9 |  | 0.127 | 0.005 | 3.5 |  | 0.158 | 0.007 | 4.4 |
| **0.3** | 0.129 | 0.006 | 6.7 |  | 0.068 | 0.003 | 3.5 |  | 0.081 | 0.004 | 4.2 |
| **0.4** | 0.101 | 0.006 | 8.8 |  | 0.039 | 0.002 | 3.4 |  | 0.046 | 0.002 | 4.0 |
| **0.5** | 0.095 | 0.002 | 12.2 |  | 0.029 | 0.001 | 3.7 |  | 0.032 | 0.002 | 4.1 |
| **0.6** | 0.103 | 0.004 | 15.0 |  | 0.030 | 0.002 | 4.3 |  | 0.028 | 0.001 | 4.0 |
| **0.7** | 0.096 | 0.005 | 14.2 |  | 0.030 | 0.002 | 4.5 |  | 0.026 | 0.002 | 3.9 |
| **0.8** | 0.122 | 0.005 | 18.0 |  | 0.057 | 0.003 | 8.4 |  | 0.029 | 0.001 | 4.3 |
| **0.9** | 0.139 | 0.004 | 19.8 |  | 0.090 | 0.005 | 12.9 |  | 0.038 | 0.002 | 5.5 |
| **1** | 0.144 | 0.007 | 22.3 |  | 0.092 | 0.004 | 14.1 |  | 0.038 | 0.002 | 5.8 |
| **<100 drugs per ADR** |
| **0.1** | 0.313 | 0.008 | 3.3 |  | 0.270 | 0.007 | 2.9 |  | 0.312 | 0.007 | 3.3 |
| **0.2** | 0.204 | 0.006 | 4.9 |  | 0.190 | 0.003 | 4.5 |  | 0.190 | 0.003 | 4.5 |
| **0.3** | 0.166 | 0.003 | 7.0 |  | 0.126 | 0.004 | 5.3 |  | 0.153 | 0.003 | 6.5 |
| **0.4** | 0.129 | 0.001 | 9.3 |  | 0.070 | 0.003 | 5.1 |  | 0.085 | 0.003 | 6.2 |
| **0.5** | 0.114 | 0.004 | 11.2 |  | 0.051 | 0.003 | 5.1 |  | 0.059 | 0.003 | 5.8 |
| **0.6** | 0.111 | 0.004 | 12.5 |  | 0.047 | 0.003 | 5.3 |  | 0.052 | 0.003 | 5.8 |
| **0.7** | 0.123 | 0.003 | 14.5 |  | 0.053 | 0.002 | 6.3 |  | 0.051 | 0.003 | 5.9 |
| **0.8** | 0.135 | 0.002 | 16.6 |  | 0.090 | 0.002 | 11.1 |  | 0.050 | 0.001 | 6.2 |
| **0.9** | 0.145 | 0.004 | 17.2 |  | 0.115 | 0.003 | 13.6 |  | 0.052 | 0.002 | 6.2 |
| **1** | 0.154 | 0.007 | 17.7 |  | 0.126 | 0.006 | 14.5 |  | 0.055 | 0.003 | 6.3 |

**Table S2.** **The results of LOOCV for chemicals of novel 3D structure using the AUPR as the accuracy measure.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **CS** |  | **ML** |  | **CCA** |
| **CutOff** | **MCC** | **STDERR** |  | **MCC** | **STDERR** |  | **MCC** | **STDERR** |
| **<12 drugs per ADR** |
| **0.1** | 0.197 | 0.009 |  | 0.030 | 0.008 |  | 0.076 | 0.004 |
| **0.2** | 0.149 | 0.007 |  | 0.057 | 0.003 |  | 0.038 | 0.004 |
| **0.3** | 0.131 | 0.006 |  | 0.014 | 0.002 |  | 0.050 | 0.002 |
| **0.4** | 0.131 | 0.006 |  | 0.014 | 0.002 |  | 0.050 | 0.002 |
| **0.5** | 0.140 | 0.010 |  | 0.014 | 0.007 |  | 0.057 | 0.005 |
| **0.6** | 0.138 | 0.010 |  | 0.007 | 0.002 |  | 0.036 | 0.003 |
| **0.7** | 0.145 | 0.004 |  | 0.023 | 0.006 |  | 0.039 | 0.005 |
| **0.8** | 0.172 | 0.011 |  | 0.050 | 0.011 |  | 0.035 | 0.004 |
| **0.9** | 0.204 | 0.011 |  | 0.087 | 0.013 |  | 0.037 | 0.005 |
| **1** | 0.204 | 0.012 |  | 0.107 | 0.005 |  | 0.041 | 0.006 |
| **<25 drugs per ADR** |
| **0.1** | 0.219 | 0.007 |  | 0.074 | 0.005 |  | 0.187 | 0.008 |
| **0.2** | 0.198 | 0.005 |  | 0.100 | 0.005 |  | 0.168 | 0.022 |
| **0.3** | 0.161 | 0.006 |  | 0.080 | 0.005 |  | 0.114 | 0.007 |
| **0.4** | 0.132 | 0.006 |  | 0.038 | 0.002 |  | 0.094 | 0.004 |
| **0.5** | 0.150 | 0.007 |  | 0.020 | 0.001 |  | 0.085 | 0.009 |
| **0.6** | 0.155 | 0.004 |  | 0.020 | 0.002 |  | 0.063 | 0.001 |
| **0.7** | 0.167 | 0.009 |  | 0.024 | 0.003 |  | 0.058 | 0.003 |
| **0.8** | 0.157 | 0.007 |  | 0.055 | 0.006 |  | 0.074 | 0.007 |
| **0.9** | 0.207 | 0.006 |  | 0.098 | 0.003 |  | 0.069 | 0.004 |
| **1** | 0.200 | 0.014 |  | 0.102 | 0.016 |  | 0.082 | 0.007 |
| **<50 drugs per ADR** |
| **0.1** | 0.245 | 0.006 |  | 0.196 | 0.005 |  | 0.185 | 0.007 |
| **0.2** | 0.224 | 0.006 |  | 0.173 | 0.005 |  | 0.224 | 0.017 |
| **0.3** | 0.184 | 0.008 |  | 0.127 | 0.006 |  | 0.163 | 0.007 |
| **0.4** | 0.149 | 0.002 |  | 0.088 | 0.002 |  | 0.129 | 0.008 |
| **0.5** | 0.174 | 0.004 |  | 0.079 | 0.001 |  | 0.099 | 0.008 |
| **0.6** | 0.161 | 0.005 |  | 0.062 | 0.002 |  | 0.086 | 0.004 |
| **0.7** | 0.171 | 0.006 |  | 0.054 | 0.002 |  | 0.085 | 0.004 |
| **0.8** | 0.163 | 0.002 |  | 0.080 | 0.001 |  | 0.082 | 0.006 |
| **0.9** | 0.200 | 0.009 |  | 0.101 | 0.005 |  | 0.095 | 0.007 |
| **1** | 0.188 | 0.006 |  | 0.114 | 0.009 |  | 0.115 | 0.006 |
| **<100 drugs per ADR** |
| **0.1** | 0.239 | 0.007 |  | 0.210 | 0.004 |  | 0.235 | 0.010 |
| **0.2** | 0.339 | 0.002 |  | 0.193 | 0.004 |  | 0.208 | 0.032 |
| **0.3** | 0.222 | 0.004 |  | 0.172 | 0.006 |  | 0.237 | 0.007 |
| **0.4** | 0.196 | 0.004 |  | 0.145 | 0.004 |  | 0.186 | 0.010 |
| **0.5** | 0.174 | 0.003 |  | 0.117 | 0.003 |  | 0.138 | 0.012 |
| **0.6** | 0.182 | 0.007 |  | 0.109 | 0.004 |  | 0.114 | 0.002 |
| **0.7** | 0.206 | 0.007 |  | 0.121 | 0.005 |  | 0.126 | 0.003 |
| **0.8** | 0.195 | 0.003 |  | 0.132 | 0.001 |  | 0.126 | 0.009 |
| **0.9** | 0.187 | 0.008 |  | 0.139 | 0.007 |  | 0.125 | 0.005 |
| **1** | 0.208 | 0.010 |  | 0.151 | 0.004 |  | 0.117 | 0.003 |

**Table S3.** **The results of LOOCV for chemicals of novel 3D structure using the MCC as the accuracy measure.**

**REFERENCES**

Duchi, J. *et al.* (2011) Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, **12**, 2121–2159.

Lim, H. *et al*. (2016) Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. *Sci. Rep.*, **6**.

Steck, H. (2010) Training and testing of recommender systems on data missing not at random. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge discovery and data mining, pages 713–722.

Yamanishi, Y. *et al.* (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.

Yao, Y. *et al*. (2014) Dual-regularized one-class collaborative filtering. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 759-768.