Вероятностный подход для задачи предсказания биологической активности ядерных рецепторов*

Володин С. Е., Попова М., Стрижов В. В.
sergei.volodin@phystech.edu
Московский физико-технический институт

Решается задача предсказания биологической активности молекул протеинов (лиганд) с рецепторами: по признакам лиганда необходимо оценить вероятность связывания этой молекулы с одним или несколькими клеточными рецепторами и построить бинарный классификатор. Экспертные знания в области биохимии и фармакологии дают основания предполагать, что факты связывания одних и тех же молекул с различными рецепторами не независимы. В данной работе предлагается модель, позволяющая строить предсказания сразу для группы рецепторов, учитывая их схожесть. В работе проводится вычислительный эксперимент на реальных данных, в ходе которого предложенная модель сравнивается с независимыми моделями в терминах нескольких функционалов качества.

Ключевые слова: *классификация, вероятность, classifier chains, multi-label, логистическая регрессия.*

Введение

Проблема предсказания биологической активности лигандов и рецепторов является актуальной задачей в области биохимии и фармакологии [1], [2], [3], [4], [5], [6]. Данная статья посвящена решению этой задачи методами машинного обучения.

Компьютерное моделирование взаимодействия молекул является распространенным методом предсказания биологической активности клеточных рецепторов [4], [1]. Однако такой способ требует знания точной структуры лиганд, которая не всегда известна. По этой причине развитие методов машинного обучения [7], позволяющих делать предсказания на основании только числовых признаков лиганд, является актуальным.

Существует два основных подхода к решению описанной задачи. В рамках первого из них для каждого клеточного рецептора строятся независимые модели. Так, например в [8], [5] применяется метод опорных векторов, в [2] и [3] — нейронные сети, а в [9] — метод k ближайших соседей. Второй подход подразумевает построение одной модели для предсказания активности группы рецепторов. Такой подход позволяет строить более сложные модели, учитывающие информацию о схожести рецепторов [6]. В [10] проведен сравнительный анализ обоих подходов.

Таким образом, данная задача решена многими способами. Тем не менее, как показывает сопоставление результатов [10], лучшим оказывается второй подход, т.е. классификаторы, учитывающие при обучении все рецепторы сразу, а не независимо друг от друга. В данном случае это означает использование нескольких классификаторов и объеднение их в «цепочку» [11], [12], [13]. Как показывает практика, обучение нескольким задачам сразу дает существенный прирост в качестве конечного алгоритма по сравнению с рассмотрением этих задач по-отдельности [14], [15], [13].

В данной работе предлагается усовершенствованный метод classifier chains [13] — вероятностная модель последовательного вывода для предсказания биологической активно-

Работа выполнена при финансовой поддержке РФФИ, проект № 00-00-00000. Научный руководитель: Стрижов В. В. Задачу поставил: Эксперт И. О. Консультант: Попова М.

сти рецепторов [16], [14]. Предложенный алгоритм относится ко второму подходу, то есть позволяет строить предсказания для группы рецепторов, а также допускает добавление новых без необходимости повторного обучения. Проведен вычислительный эксперимент на реальных данных, в котором набор независимых моделей сравнивался с моделью последовательного вывода. Построенные модели сравнивались по нескольким критериям качества.

Постановка задачи классификации

Задана выборка $\mathfrak{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in\mathcal{L}}, \mathcal{L} = \{1, \ldots, m\}$ — m пар объект-ответ. Каждый из объектов $\mathbf{x}_i \in \mathbb{R}^n$ — вектор действительных чисел. Объект может принадлежать каждому из l, что представляется вектором ответов $\mathbf{y}_i \in \{0, 1, \square\}^l$, 1 означает принадлежность классу, а $\square$ означает пропуск в данных. Выборка разбита на обучающую и контрольную: $\mathfrak{D} = \mathfrak{L} \sqcup \mathfrak{T}$

Определяются $\mathbf{X}, \mathbf{Y}$ — случайные величины. Считается, что между классами есть зависимости:

$$P(\mathbf{Y}|\mathbf{X}) \neq \prod_{j=1}^{l} P(y_j|\mathbf{X})$$

Вводится предположение, что условное распределение $P(\mathbf{X}|\mathbf{Y})$ принадлежит семейству экспоненциальных распределений.

Моделью классификации называется функция $\mathrm{f}\colon \mathbf{W} \times \mathbf{X} \times \mathbf{Y} \to [0, 1]$, где $\mathbf{W}$ — множество параметров, $\mathbf{w} \in \mathbf{W}$ — вектор параметров модели. Значение f — апостериорая вероятность:

$$\mathrm{f}(\mathbf{w}, \mathbf{x}, \mathbf{y}) = P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}; \mathbf{w})$$

Функция потерь для значения параметра $\mathbf{w}$ и подвыборки $\mathcal{Z}$ определяется через функцию правдоподобия модельного распределения:

$$Q(\mathrm{f}|\mathbf{w}, \mathcal{Z}) = -\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{Z}} \log \mathrm{f}(\mathbf{w}, \mathbf{x}, \mathbf{y}) P(\mathbf{X} = \mathbf{x})$$

Требуется найти вектор параметров $\mathbf{w}^* \in \mathbf{W}$, минимизирующий Q на обучающей выборке $\mathfrak{L}$:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbf{W}} Q(\mathrm{f}|\mathbf{w}, \mathfrak{L})$$

Поскольку выборка содержит пропуски, разбиения должны быть построены таким образом, чтобы в каждой подвыборке было достаточное количество объектов с известным значением каждого признака.

В качестве дополнительного критерия качества модели используются значения функционала AUC для каждого класса j на контрольной выборке $\mathfrak{T}$ при 5 различных разбиениях.

Базовый вычислительный эксперимент

Целью эксперимента является получение характеристик простого алгоритма для дальнейшего сравнения его с предлагаемым в статье. Базовый алгоритм использует подход Binary Relevance [14], в котором зависимости между классами не учитываются. Таким образом, алгоритм представляет собой l независимых логистических регрессий, по одному классификатору для каждого класса.

Таблица 1. Количество связывающихся с рецепторами лигандов

Рецептор	Неизвестно	Не связывается	Связывается
NR-AhR	**3413** (40%)	**4503** (52%)	**597** (7%)
NR-AR-LBD	**3213** (37%)	**5129** (60%)	**171** (2%)
NR-AR	**2904** (34%)	**5398** (63%)	**211** (2%)
SR-MMP	**3925** (46%)	**3870** (45%)	**718** (8%)
NR-ER	**3746** (44%)	**4232** (49%)	**535** (6%)
SR-HSE	**3309** (38%)	**4961** (58%)	**243** (2%)
SR-p53	**3174** (37%)	**5029** (59%)	**310** (3%)
NR-PPAR-gamma	**3393** (39%)	**4987** (58%)	**133** (1%)
SR-ARE	**3791** (44%)	**4029** (47%)	**693** (8%)
NR-Aromatase	**4544** (53%)	**3835** (45%)	**134** (1%)
SR-ATAD5	**2951** (34%)	**5360** (62%)	**202** (2%)
NR-ER-LBD	**3107** (36%)	**5168** (60%)	**238** (2%)

Эксперимент проведен на реальных данных, имеющих двойное происхождение. Объектами являются лиганды, их признаки $\mathbf{x}_i$ смоделированы при помощи специальной программы. Ответы $\mathbf{y}_i = (y_{i1}, ..., y_{il})$ являются результатами биохимических экспериментов, показывающих, связывается ли данный лиганд с рецептором j. Пропуск в ответах означает, что эксперимент либо не был проведен, либо не позволяет с достаточной уверенностью говорить о каком-либо результате. Каждый объект имеет 165 признаков. Признаки являются химическими параметрами молекулы. В выборке содержится 8513 объектов, количество объектов с измеренным ответом j составляет около половины. В таблице 1 указано точное распределение ответов по классам.

Для определения эффективности данного метода вычисляются значения функционала AUC для каждого из разбиений $\mathfrak{D} = \mathfrak{L} \sqcup \mathfrak{T}$ на тестовую и контрольную выборку. Разбиения выполнены по методу k-fold, где $k = 5$. Вычисляется среднее значение AUC, а также стандартное отклонение.

На графиках (1, 2) показаны ROC-кривые классов для одного из разбиений, а также значение функционала AUC.

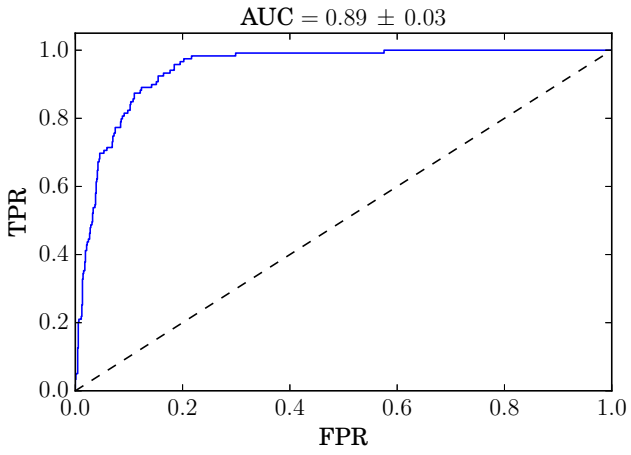
В таблице 2 приведено сравнение метода Binary Relevance с результатами из [17], для получения которых использовались те же данные и способ разбиения, что и в данной работе.

Сравнение результатов показывает, что простой алгоритм уступает в качестве классификации методу Random Forest. Для некоторых рецепторов эта разница значительна.

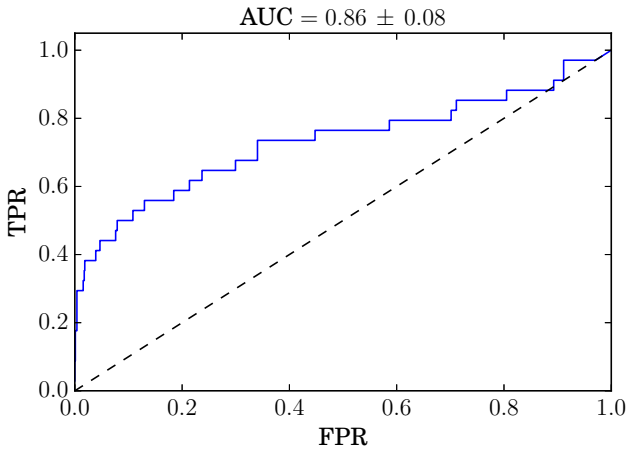
Описание алгоритма

Вычислительный эксперимент

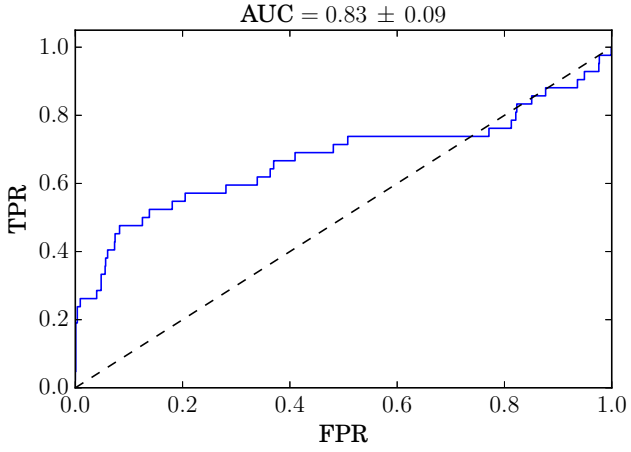
Заключение



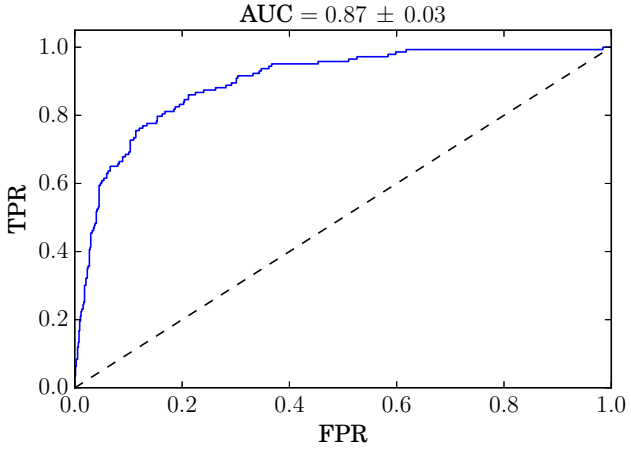
(a) Рецептор NR-AhR



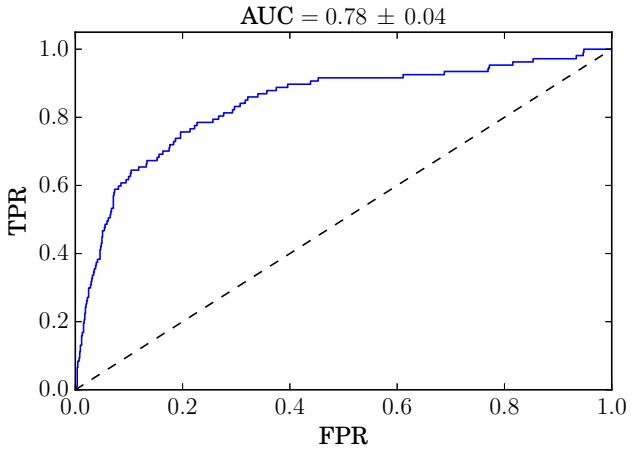
(b) Рецептор NR-AR-LBD



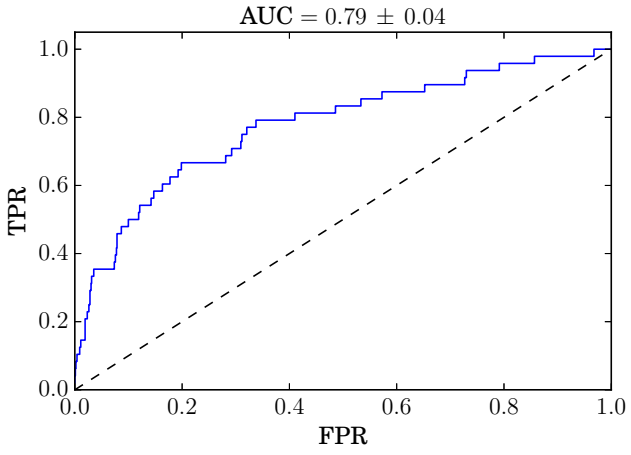
(c) Рецептор NR-AR



(d) Рецептор SR-MMP

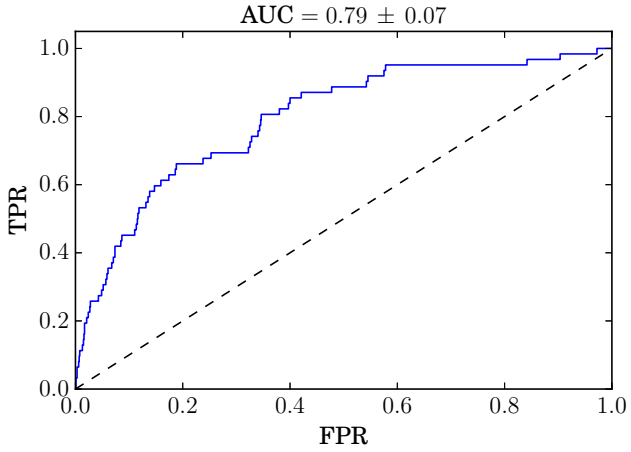


(e) Рецептор NR-ER

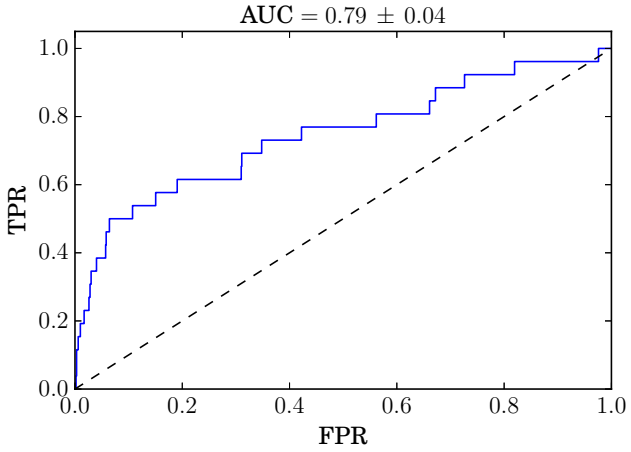


(f) Рецептор SR-HSE

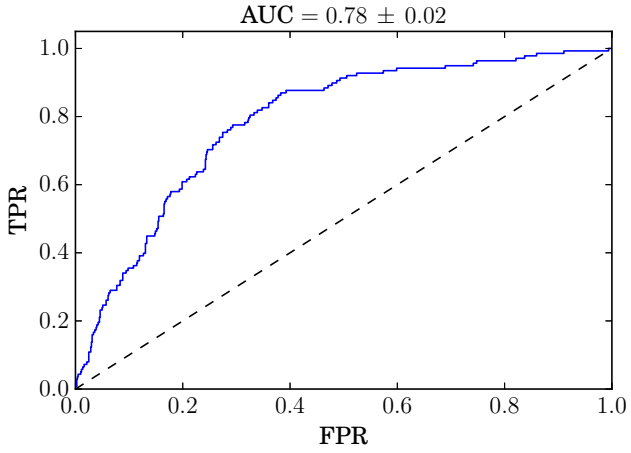
Рис. 1. ROC-кривая и значения функционала AUC для классов 1-6, метод Binary Relevance



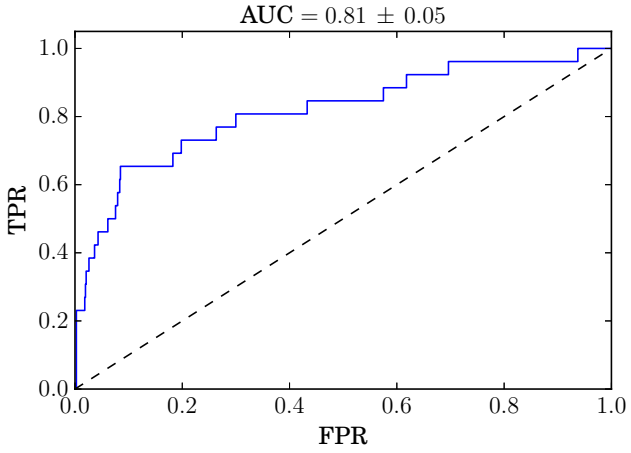
(a) Рецептор SR-p53



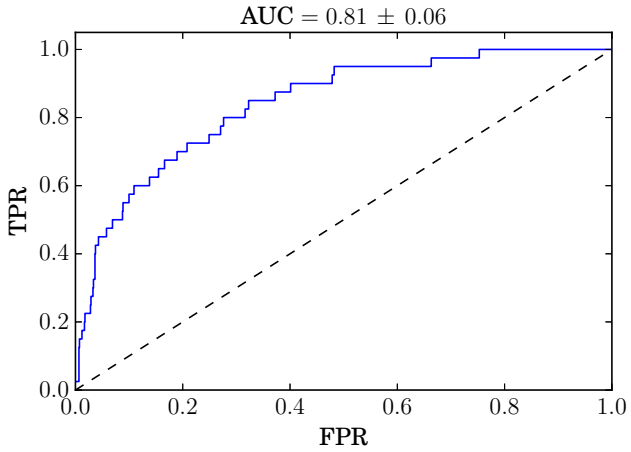
(b) Рецептор NR-PPAR-gamma



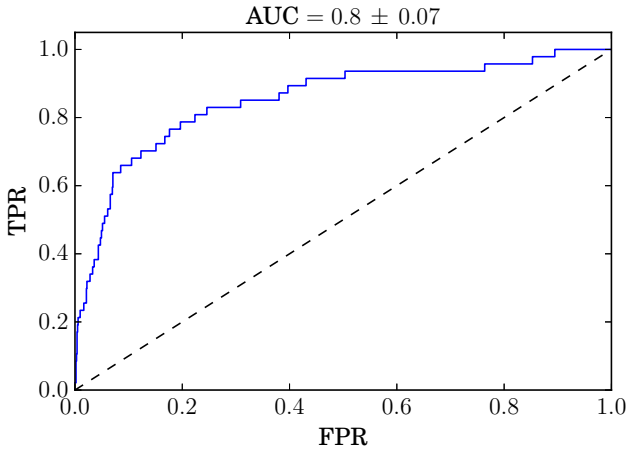
(c) Рецептор SR-ARE



(d) Рецептор NR-Aromatase



(e) Рецептор SR-ATAD5



(f) Рецептор NR-ER-LBD

Рис. 2. ROC-кривая и значения функционала AUC для классов 6-12, метод Binary Relevance

Таблица 2. Значение AUC для различных рецепторов и моделей классификации

Рецептор	Binary Relevance	Random Forest [17]
NR-AhR	**0.83** $\pm$ 0.03	**0.93**
NR-AR-LBD	**0.86** $\pm$ 0.08	**0.88**
NR-AR	**0.83** $\pm$ 0.09	**0.83**
SR-MMP	**0.87** $\pm$ 0.03	**0.95**
NR-ER	**0.78** $\pm$ 0.04	**0.81**
SR-HSE	**0.79** $\pm$ 0.04	**0.86**
SR-p53	**0.79** $\pm$ 0.07	**0.88**
NR-PPAR-gamma	**0.79** $\pm$ 0.04	**0.86**
SR-ARE	**0.78** $\pm$ 0.02	**0.84**
NR-Aromatase	**0.81** $\pm$ 0.05	**0.84**
SR-ATAD5	**0.81** $\pm$ 0.06	**0.83**
NR-ER-LBD	**0.80** $\pm$ 0.07	**0.83**

Литература

[1] R. DVORSKÝ V HORŇÁK and E. ŠTURDÍK. Receptor-ligand interaction and molecular modelling.

[2] Tong Q Xie XQ Myint KZ, Wang L. Molecular fingerprint-based artificial neural networks qsar for ligand biological activity predictions. *Molecular Pharmaceutics*, 2012.

[3] Xie XQ Myint KZ. Ligand biological activity predictions using fingerprint-based artificial neural networks (fann-qsar). *Methods Mol. Biol.*, 2015.

[4] Bonnie Berger Vinay Pulim, Jadwiga Bienkowska. Lthreader: Prediction of extracellular ligand–receptor interactions in cytokines using localized threading. *Protein Science*, 2008.

[5] Changhong Zhou Wenjun Zhang Zhengjun Cheng, Yuntao Zhang and Shibo Gao. Classification of 5-ht1a receptor ligands on the basis of their binding affinities by using pso-adaboost-svm.

[6] Laurent Jacob and Jean-Philippe Vert. Protein–ligand interaction prediction: an improved chemogenomics approach. *BIOINFORMATICS*, 2008.

[7] Peter Willett. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 1998.

[8] Yusuke Komiyama et al. Masayuki Yarimizu, Cao Wei. Tyrosine kinase ligand-receptor pair prediction by using support vector machine. *Advances in Bioinformatics*, 2015.

[9] Nagamani Sukumar Curt Breneman Scott Oloff†, Shuxing Zhang and Alexander Tropsha. Chemometric analysis of ligand receptor complementarity: Identifying complementary ligands based on receptor information (colibri). *J. Chem. Inf. Model.*, 2006.

[10] M. Popova. Feature selection and multi-task prediction of biological activity for nuclear receptors. 11(1):111–112, 2015.

[11] Jose Barranqueroa José Ramón Quevedoa Juan José del Coza Eyke Hüllermeierb Elena Montañesa, Robin Sengeb. Dependent binary relevance models for multi-label classification. *Pattern Recognition*, 2013.

[12] Ivor W. Tsang Weiwei Liu. On the optimality of classifier chain for multi-label classification.

[13] Geoff Holmes Eibe Frank Jesse Read, Bernhard Pfahringer. Classifier chains for multi-label classification.

[14] Eyke H.0 Krzysztof Dembczynski, Weiwei Cheng. Bayes optimal multilabel classification via probabilistic classifier chains. 2010.

[15] Haytham Elghazel Maxime Gasse, Alex Aussem. On the optimality of multi-label classification under subset zero-one loss for distributions satisfying the composition property. 2015.

[16] Eduardo F. Morales Pablo Hernandez-Leal Julio H. Zaragoza Pedro Larrañaga L. Enrique Sucar, Concha Bielza. Multi-label classification with bayesian network-based chain classifiers.

[17] Olexandr Isayev Sherif Farag Stephen J. Capuzzi, Regina Politi and Alexander Tropsha. Qsar modeling of tox21 challenge stress response and nuclear receptor signaling toxicity assays.