2014-01-10

Getting started with Bioclipse Modeling

Introduction

Bioclipse Modeling is the flagship product of Genetta Soft. The solution extends the Bioclipse Decision Support tool with the capability to produce accurate and robust predictive models starting from a collection of chemical structures with an associated response value (such as an interaction or toxicity measurement/classification).

Installation

Download one of the pre-built packages from www.genettasoft.com/downloads. Unzip the downloaded archive onto your local computer, and start Bioclipse by double-clicking on the icon in the Bioclipse folder.

Bioclipse Decision Support

Bioclipse Decision Support is an open source solution that is capable of executing multiple predictive models simultaneously on chemical structures, and deliver timely, interpretable results. If you use Bioclipse Decision Support in your research, please cite:

> O. Spjuth, L. Carlsson, E. Ahlberg-Helgee, M. Eklund, and S. Boyer
> Integrated decision support for chemical liability
> *J. Chem. Inf. Model.*, 2011, 51 (8), pp 1840-1847

The Decision Support Perspective
A Decision Support Perspective is available to arrange the Bioclipse workbench for predicting chemical liabilities. From the main menu, choose **Window > Open Perspective** and then select the **Decision Support** option.

Predicting for individual molecules: The Decision Support View
Open the view **Decision Support**, available from the menu **Window > Show view > other...** and select **Decision Support**. Now, open a chemical structure with 2D coordinates in the chemical editor by double-clicking on e.g. a mol/cml file in the **Bioclipse Navigator**. Click the **Run** button in the local toolbar of the **Decision Support** view. If there are matches, you can select them in the Decision Support view and get more detailed decision support, for example a highlighted

substructure or the full structure of an identical hit in an external database. Results may look like the figures below.

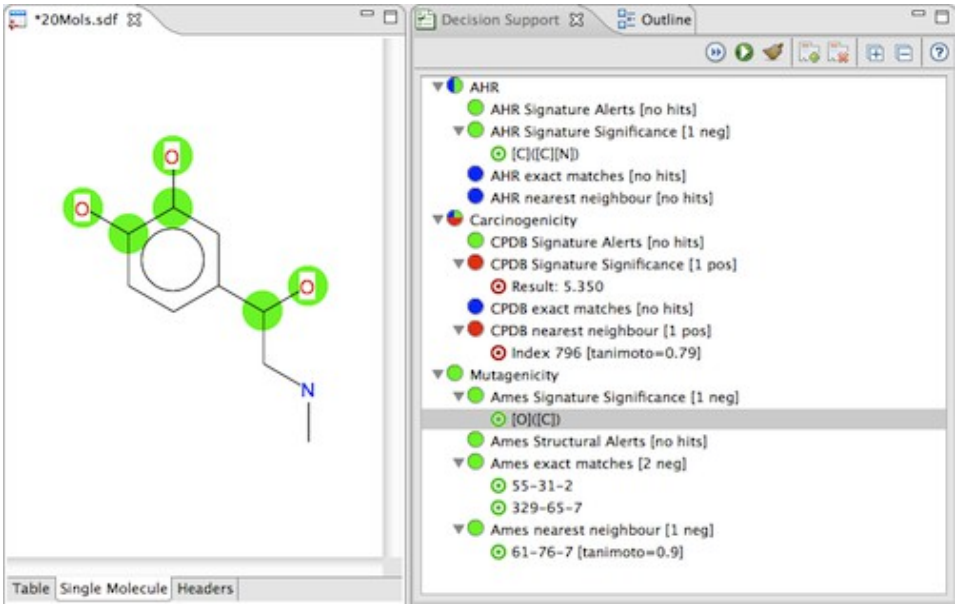

Figure 1: Models resulting in substructure matches can be highlighted in the original chemical structure.

You can make changes in the chemical editor and re-run the models to get an updated result. Note the option in the Decision Support View local toolbar to auto-rerun on a chemical edit. This way it is possible to, in real time, make changes and see how these changes are predicted, allowing for trying different hypothesis.

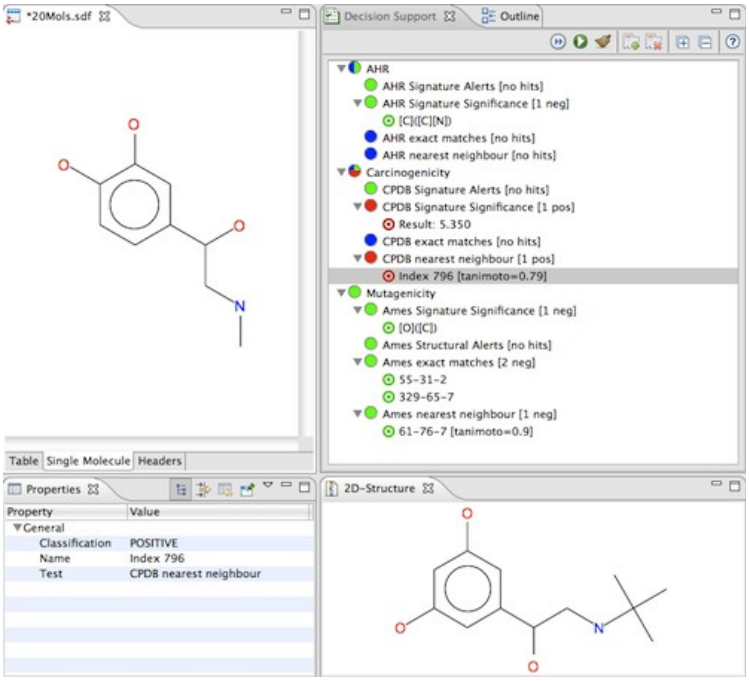

Figure 2: Models resulting in external matches can be visualized in the 2D-Structure View.

Bioclipse Modeling

Bioclipse Modeling allows you to build predictive models from a collection of molecules, available in SD-Files, with a property for each molecule describing the endpoint to be modeled (such as interaction, toxicity, etc). Bioclipse Modeling supports both regression models as well as classification models.

To start Bioclipse Modeling, right-click an SD-file and select from the popup meny: Bioclipse Modeling > New Model... This opens the Bioclipse Modeling wizard.

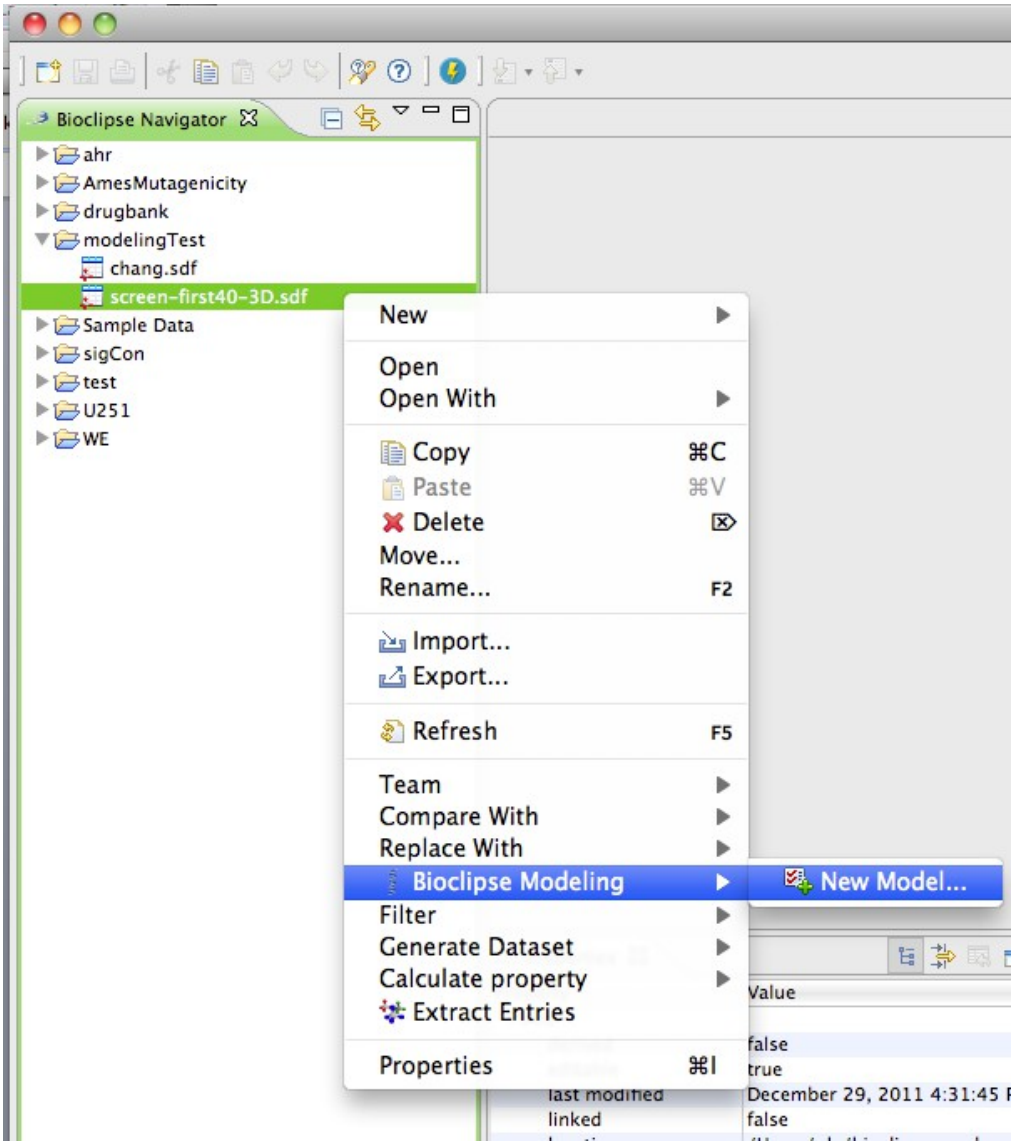

Figure 3: Right-click an SD-file with chemical structures and select Bioclipse Modeling > New Model... from the context menu to start Bioclipse Modeling.

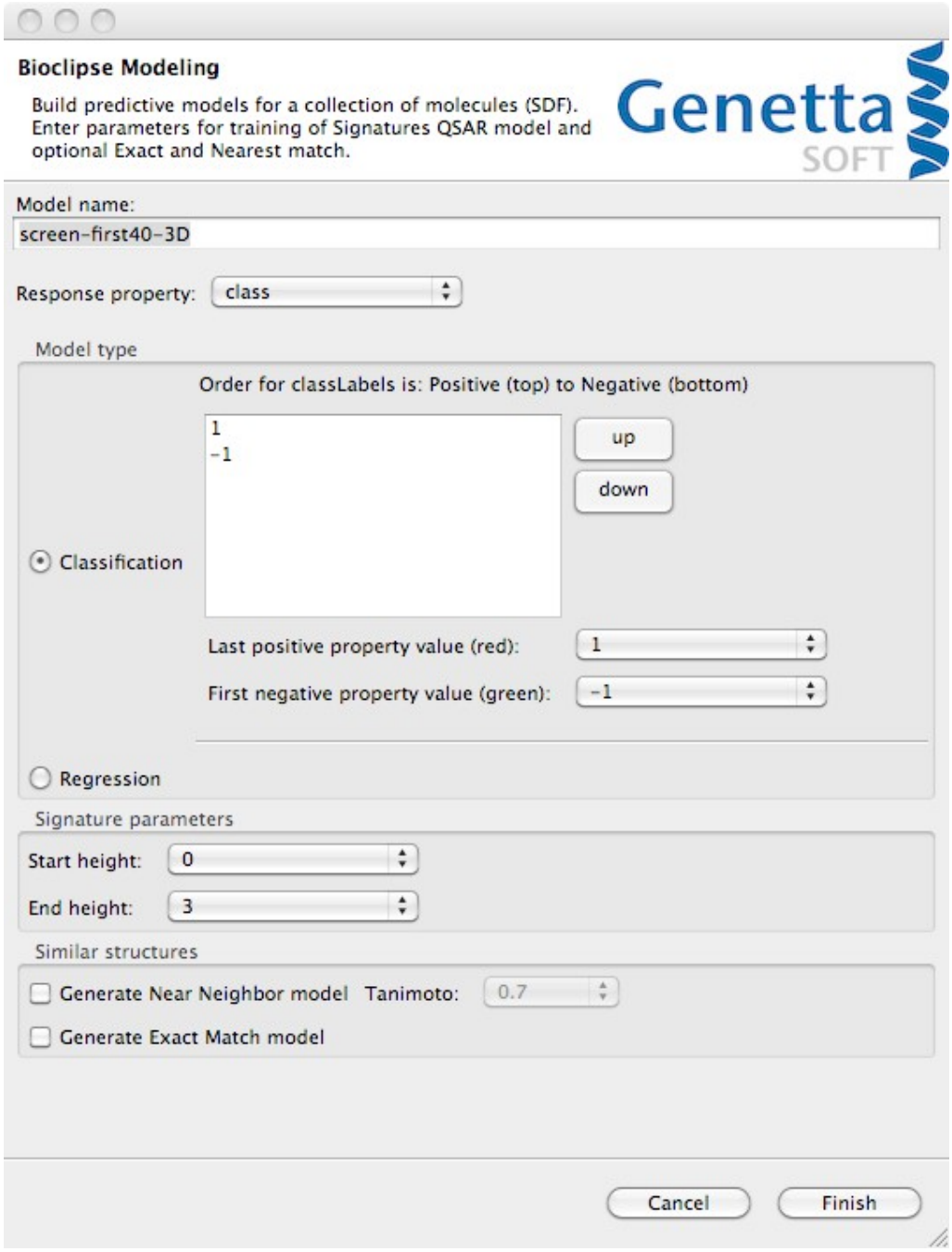

Figure 4: Bioclipse Modeling Wizard with a Classification model selected; the property "class" in the SD-file has the values 1 or -1 for every molecule.

Parameter descriptions

Model name: Here you can enter a name for the model to be built.

Response property: This is the property for each molecule that will be modeled. Select the desired property from the combo-box.

Model Type: If the property to be modeled has discrete values (such as 2 classes e.g. mutagen/nonmutagen), then select the **Classification** radio button (this is the default). You may arrange the class order by selecting them in the frame and pressing the buttons up or down. Use the combo boxes "Last positive property value" and "First negative property value" to assign coloring for the resulting model interpretation. Read more about this below. See Figure 4 for how a classification example looks like.

If the property to be modeled has continuous values, such as a measured interaction (e.g. IC_{50}) then select the **Regression** radio button. See Figure 5 for how this may look like.

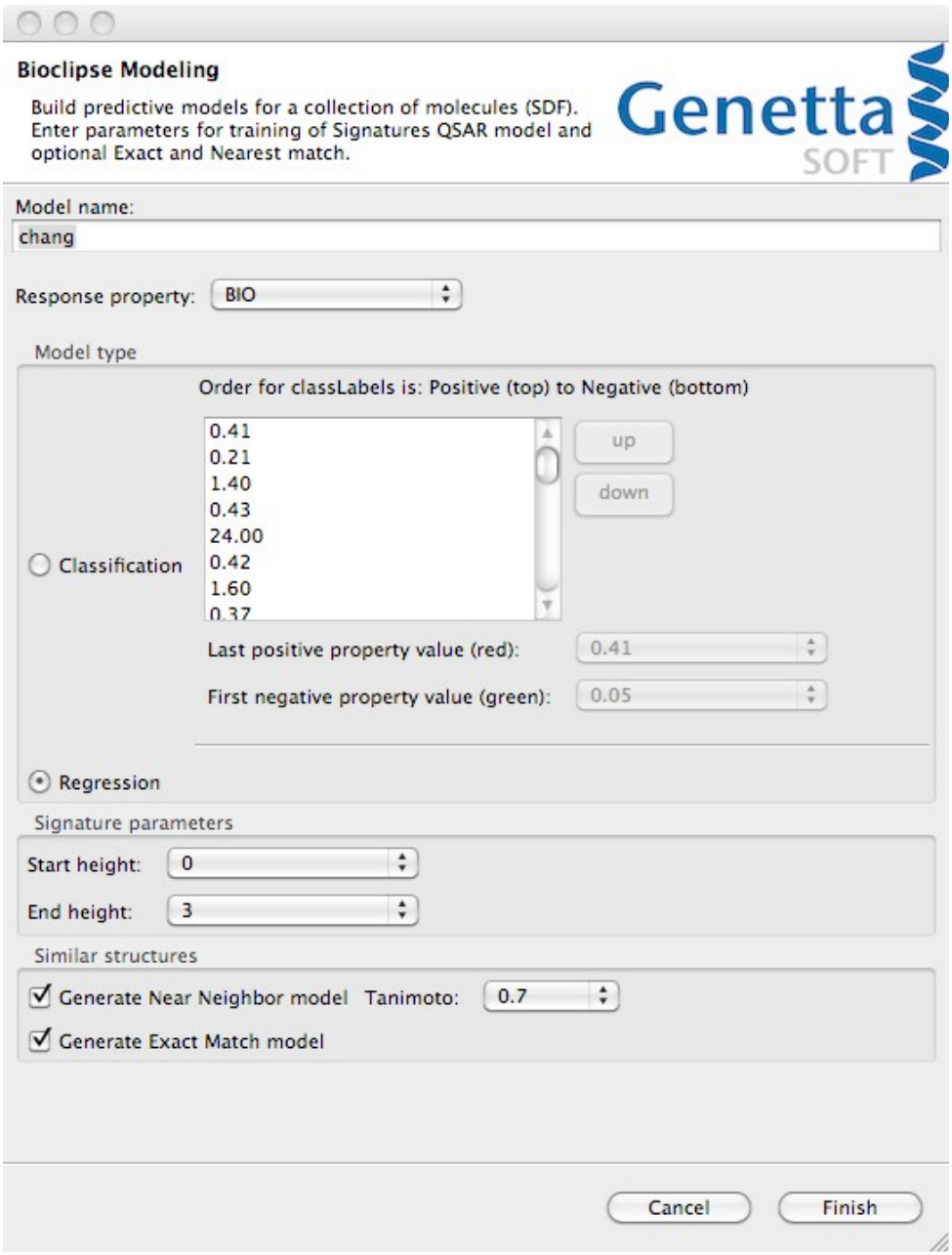

Figure 5: Regressiom model type selected in the Bioclipse Modeling Wizard.

Signature parameters: Here you can select the starting and ending height of the chemical descriptors – Signatures. These are robust descriptors that has proven useful in modeling as they deliver accurate models with results that can be interpreted as substructures. Read more on Signatures in Appendix 1.

Similar structures: Here you can choose to generate two additional models; one for near neighbours (based on an 1024 bit fingerprint) where you also need to input a tanimoto measure, above which a compound is interpreted as a near neighbor. Exact Match will display a hit if a query molecule is identical to a chemical structure in your SD-file. Read more on implementations in Appendix 1.

Model building

Click Finish to start model building, and Bioclipse will build the selected models. A progress monitor will be displayed, and when finished a dialog will appear with statistics on the QSAR model built (see Figure 6 for an example with a classification model). The optimum value is R^2 for classification model, and RMSE for regression models (see Appendix 1).

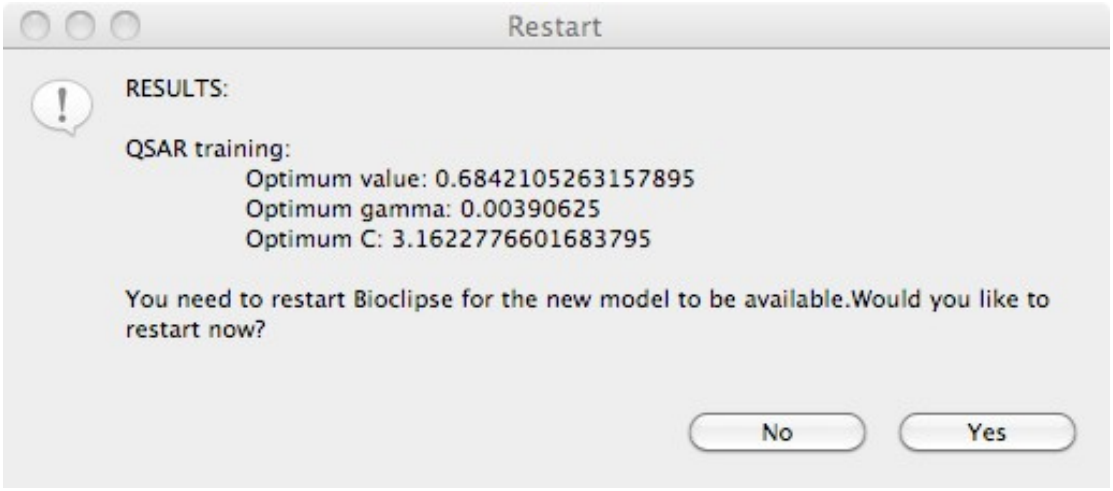

Figure 5: Result dialog for a classification model.

Using the models

A restart of Bioclipse Modeling is required, and after this the new models will be available in the Decision Support View. Clicking on a QSAR model will show training dataset information, including number of observations (signatures) variables (molecules), as well as model information including parameters for the SVM model (see Figure 6).

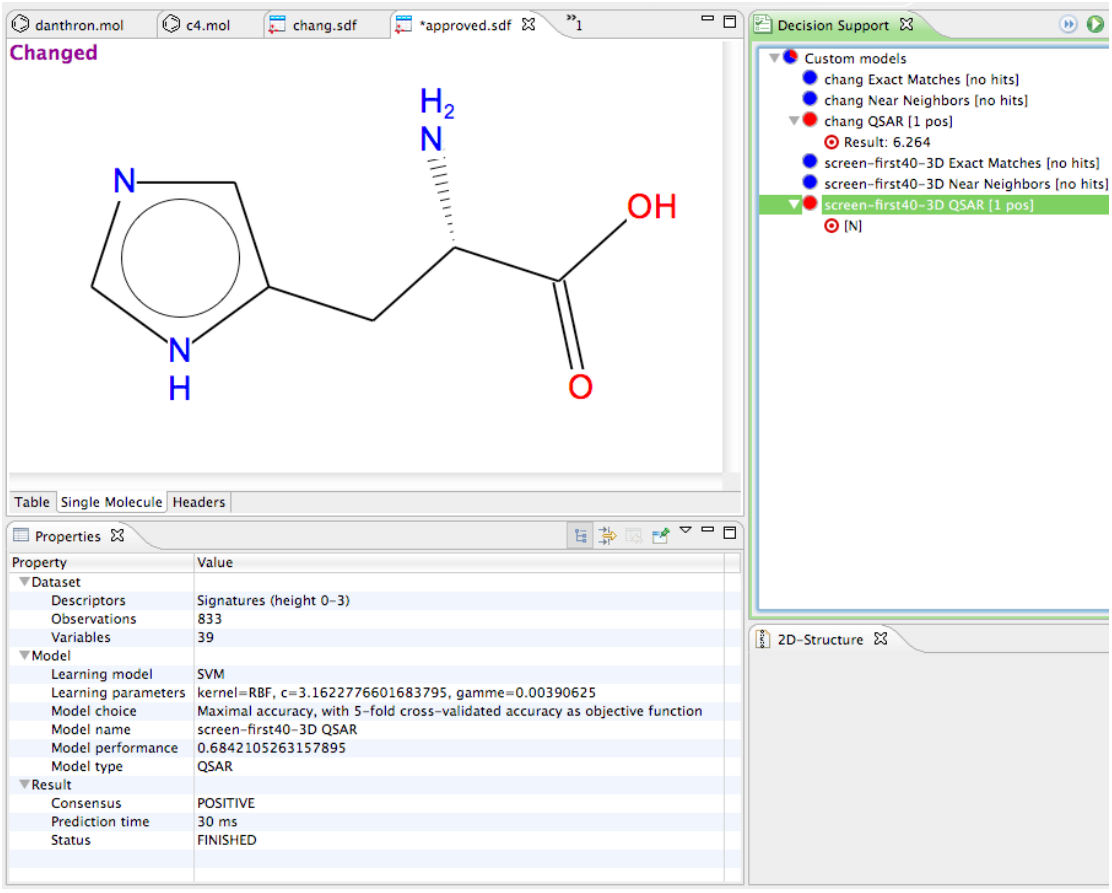

Figure 6: Clicking a QSAR model shows information about the dataset and the resulting model.

Open a chemical structure and press the **Run** button in the Decision Support View toolbar, and when models are finished you can select them to view details on the result and highlight substructures for the QSAR models.

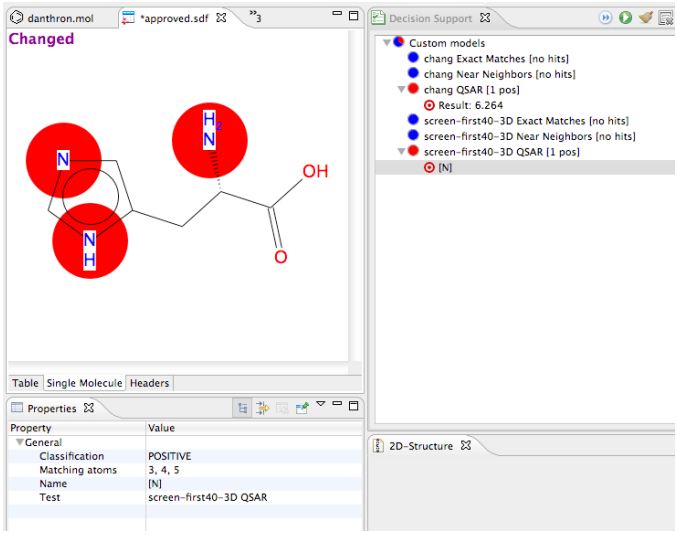

Figure 7: Results from a QSAR models highlighted. The atoms colored in red contributed the most to a positive prediction.

Appendix 1: Methods

Representation of chemistry

Bioclipse Modeling makes use of Faulon Signatures for describing chemistry, see the following article for a more detailed description.

Churchwell C. J., Rintoul M. D, Martin S., Visco D., Kotu, A., Larson R. S., Sillerud L. O. Brown D. C., Faulon J.L.
The Signature Molecular Descriptor. 3. Inverse Quantitative Structure-Activity Relationship of ICAM-1 Inhibitory Peptides
J. Molecular Graphics & Modelling, 22, 263-273, 2004.

Modeling method

Bioclipse Modeling applies a Support Vector Machine (SVM) to build predictive models, with an RBF kernel. A grid search to find optimal values for the parameters C and GAMMA is applied. For an introduction to SVM we refer to the introductory guide: *A Practical Guide to Support Vector Classifcation*, available from: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Model validation

For model validation we apply a 5-fold cross-validation and report R^2 in case of classification, and RMSE in case of regression.

Near Neighbors

Bioclipse Modeling uses 1024 bits CDK Fingerprints to represent chemical structures and the tanimoto distance for similarity searches.

Exact match

Bioclipse Modeling uses Molecular Signatures, which are Faulon Signatures of enough height to encompass the entire structure, in order to query for exact matches in the provided training set. This is very similar to the InChI chemical identifier but with increased performance.