Nanoinformatics Roadmap & Modeling

Roadmap Status

Definition of Database

Treatment of materials modeling and QSARs

Agenda

- Roadmap Status
  - Draft posted in December
  - Comments received through mid-January
  - Definitions; Ontology; Chemoinformatics & Materials Modeling Sections revised
  - Waiting for lead authors to respond
- Overview of general comments
- Details of Ontology & Modeling open items

General Comments

- SWENanoSafe; 5-6 colleagues
  - Major change in format, e.g. Introduction is Section 4 and not at beginning or merge 9 & 10 **(too late)**
  - Need better 1st paragraphs in Executive; Milestones; and other locations **(addressed)**
  - Difficult to find the Roadmap **(Retitled Section 12)**
- EU Materials Modeling Group (meeting notes)
  - Add Validation chapter; expand on curation **(too late)**
  - Intend one or many databases? **(written as many)**
  - Planned ontology for Materials Modeling **(mentioned)**
  - Distinguish data-derived descriptors from *ab initio* ones **(currently under discussion)**

Yoram Cohen Comment

- Current definition of database
  - Structured electronic dataset
- The term …..should be "Structured Database" Note that a database can also be "A database of unstructured datasets". A database of structured dataset as the definition implies is restricted to typical relational databases that contain fixed fields and records. This is the "old" school approach and with the advent of Big Data, there are various systems that utilize approaches more suitable for unstructured datasets (e.g., Graph Database)

Issues

- The IT literature invariably uses ‘structure’
  - Structured storage
  - Database system is structured in a formal language
  - Hierarchical structures
  - Graph structures
  - Schema is a complete description of the structure of a database
- Structure is as confusing in IT as it is in nanoEHS, e.g. nanostructures
- Use ISO definition

Current & Proposed Definitions

- Was
  - Structured electronic dataset
- Now (ISO/IEC 2382:2015)
  - collection of data organized according to a conceptual structure describing the characteristics of these data and the relationships among their corresponding entities, supporting one or more application areas

- Recall and precision (library science) are the minimum requirements for ‘organized’

Questions

- Do modelers require an ontology or do they require a dataset? **Yes & No**
- Does materials modeling require a materials ontology with which to model? **Yes**
- Is QSAR modeling different than materials modeling relative to datasets found in databases? **Yes & No**
- Is there a core set of eNanoMapper (& NPO) terms that should be consistent across any nanoEHS ontology? **Yes & No**
- Is compatibility with ISA-TAB essential? **No**

Terms-to-Models Corridor

- Chemical Structure (basis of OECD documents)
- Composition is a mixture of ‘multi-structural substances’ at OECD
- Molecular identity: TSCA basis of chemical substance
- Molecular Structure (same as chemical structure?)
- Categories & Analogs (basis of EPA review of PMNs)
- Grouping (emphasized at ECHA)
- QSPR; QSAR; read-across (case-by-case acceptance)

FK: molecular structure implicit to all terms (ethanol), but particles have stoichiometric compositions without molecular structures, e.g. SiO2 or **water ≠ H_2O**

Model Descriptors

- Is a correlation step always necessary ?
- Material modelers believe *ab initio* (fundamental) models do not require correlation to biological data
- They combine a physics equation with a material relation, but may need to use a solver or do post-processing
- There is no recognized cause & effect for AOPs or cellular assays; there are some plausible modes of action (MoA)
- Descriptors carry assumptions, as do solvers, and therefore correlation step is necessary

Model Example

$$\mathbf{y} = b_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \ldots + b_n\mathbf{x}_n$$

$$\log\left(1/EC_{50}\right) = 2.59 - 0.50 \cdot \Delta H_{Me+}$$

- This is an Arrhenius eqn. for 1st order kinetics or a reaction with an activation energy
- EC_{50} is half the biological reaction rate, like half-life, $t_{1/2}$
- 10 of 12 structural descriptors were energy
- Winnowed descriptors in survival of the fittest manner
- Exponential function ‘favors’ energy as answer, but activation energy ≠ equilibrium enthalpy (& entropy?)

Model Validation by Burello

- Published QSARs do not comply fully with OECD

1. Consider all potentially relevant descriptors
2. Descriptors should reflect measured properties
3. Need descriptor's statistical relevance to prediction
4. Descriptors with p-values >0.05 should be disregarded to avoid wrong mechanistic interpretations
5. Low p-values better as changes can be related to changes in response variable

Model Conclusion

- Model validation by modelers (internal consistency & robustness) ≠ model validation by regulators
- For regulator, the model will likely need:
  - Plausible reasons for descriptor selection (mechanism or *in vitro* test results);
  - More than one descriptor, probably 3;
  - Internal validation will be matching a measured property to the model's prediction of properties associated with the descriptors;
  - EHS validation will be the fit to biological data; and
  - Explanation for any intercepts, etc. derived from the math
- Models are virtual particles having a limited set of properties, those sufficient to account for toxicity

Terms-to-Models Corridor

- Recent conversations prompted by Roadmap
  - Standardization is knowledge codification (avoid lock-in)
  - AOPs are knowledge organization (not cause & effect)
  - CEIN’s dynamic classification is not an ontology, but allows for recall and precision.
- Anchoring the nanoEHS database to the regulatory framework means mapping the science to the corridor from ‘chemical structure’ to QSAR
- Anchoring the nanoEHS database is a multi-disciplinary effort, the term being defined becomes a boundary object, e.g. structure.