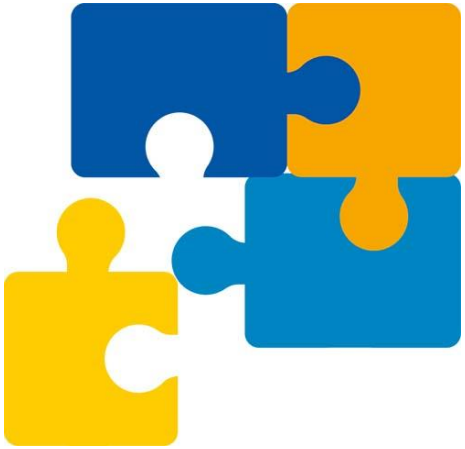
Grouping of substances in screening

Webinar: how are substances shortlisted and manually screened?

1 February 2018

Giovanni Bernasconi
European Chemicals Agency

Content



- Substance grouping
  - Why shifting towards screening of groups
  - Methodology used to build groups
- How grouping was applied in round5 of screening
  - Type of shortlisted groups
  - MSCAs screening and tips for Registrants
- Conclusions

Substance grouping

Why shifting towards groups?

- Common screening is constantly evolving and improving
  - First rounds were based on IT screening scenarios and shortlisting of individual substances
  - From rounds 2 and 3 similarity within shortlisted and CoRAP substances was flagged
  - Since round 4 similarity and grouping has been used to build the shortlist, i.e. analogues to substances of concern ('seeds') were identified and included in the shortlist

- Looking at substances in isolation is not the optimal approach and grouping is essentially unavoidable
  - For most identified substances, action is ongoing on it or a relative

Benefits of working with groups

- By pooling together all hazard information for related substances it may be possible to conclude manual screening despite data gaps for individual substances
- By looking at the whole group, including substances for which information generation is being considered or on going, it may be easier to fine tune our regulatory actions
  - Target the right substance at the right time
- Consistency in how related substances are treated
- Fairness to industry and better informed substitutions

Methodology used to form groups

Grouping similar substances: how?

1. structural information

- substance identity information in IUCLID
- external sources to convert names and numerical identifiers into structures

2. read across & category information

- test material identifiers in endpoint study records (read-across information)
- category objects in registration dossiers
- external sources with category information

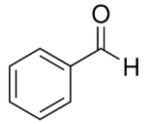
➢ *Future developments: uses, structural alerts, MoA, metabolism prediction…*

➢ Finding related substances at the stage of screening is not the same as fulfilling the criteria of Annex XI, 1.5 of REACH (grouping and read-across)

Grouping by structural similarity – similarity index

- Molecular structures are "broken down" to functional groups taking into account connectivity up to a given distance ("chemical vocabulary")
- Every molecular structure is converted into a binary vector (vector with zeroes and ones)

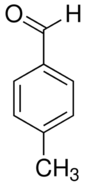
substance *i*
(registered)



$\text{bitvector } \boldsymbol{X}_i \in R^L$

11101010…00000001

substance *j*
(in the reference pool)



$\text{bitvector } \boldsymbol{X}_j \in R^L$

11101000…00000011

several aspects need to be considered, such as the length and density of the fingerprint, and the nature of the chemical vocabulary

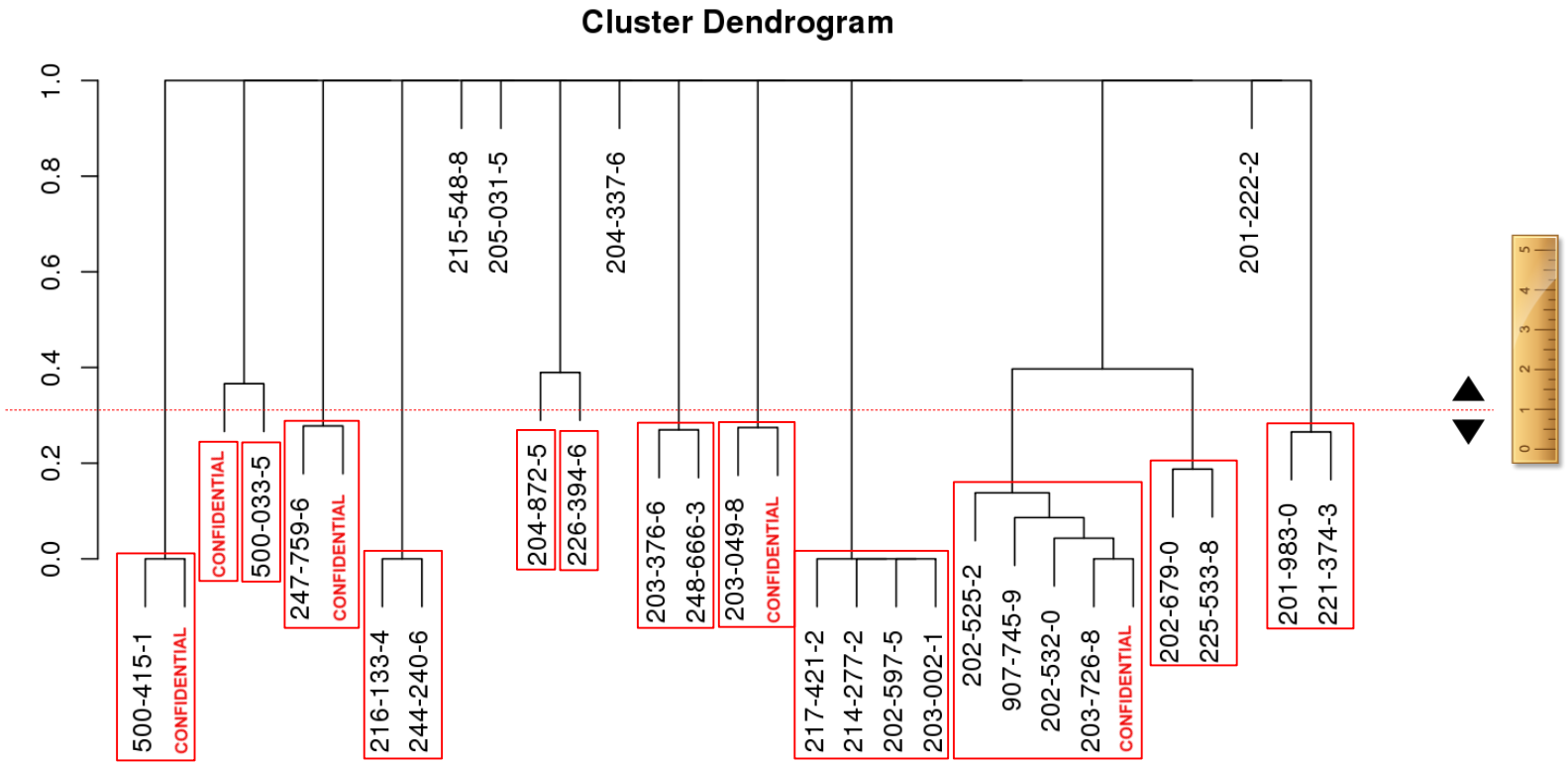
- We compute the distance using a distance function (typically Tanimoto)

$$d_{ij} = T_{ij} = 1 - \frac{\boldsymbol{X}_i \cdot \boldsymbol{X}_j}{|\boldsymbol{X}_i|^2 + |\boldsymbol{X}_j|^2 - \boldsymbol{X}_i \cdot \boldsymbol{X}_j} \in [0,1]$$

distance = 0 means identical structures

distance = 1 means completely different structures

Grouping by structural similarity – dendrogram



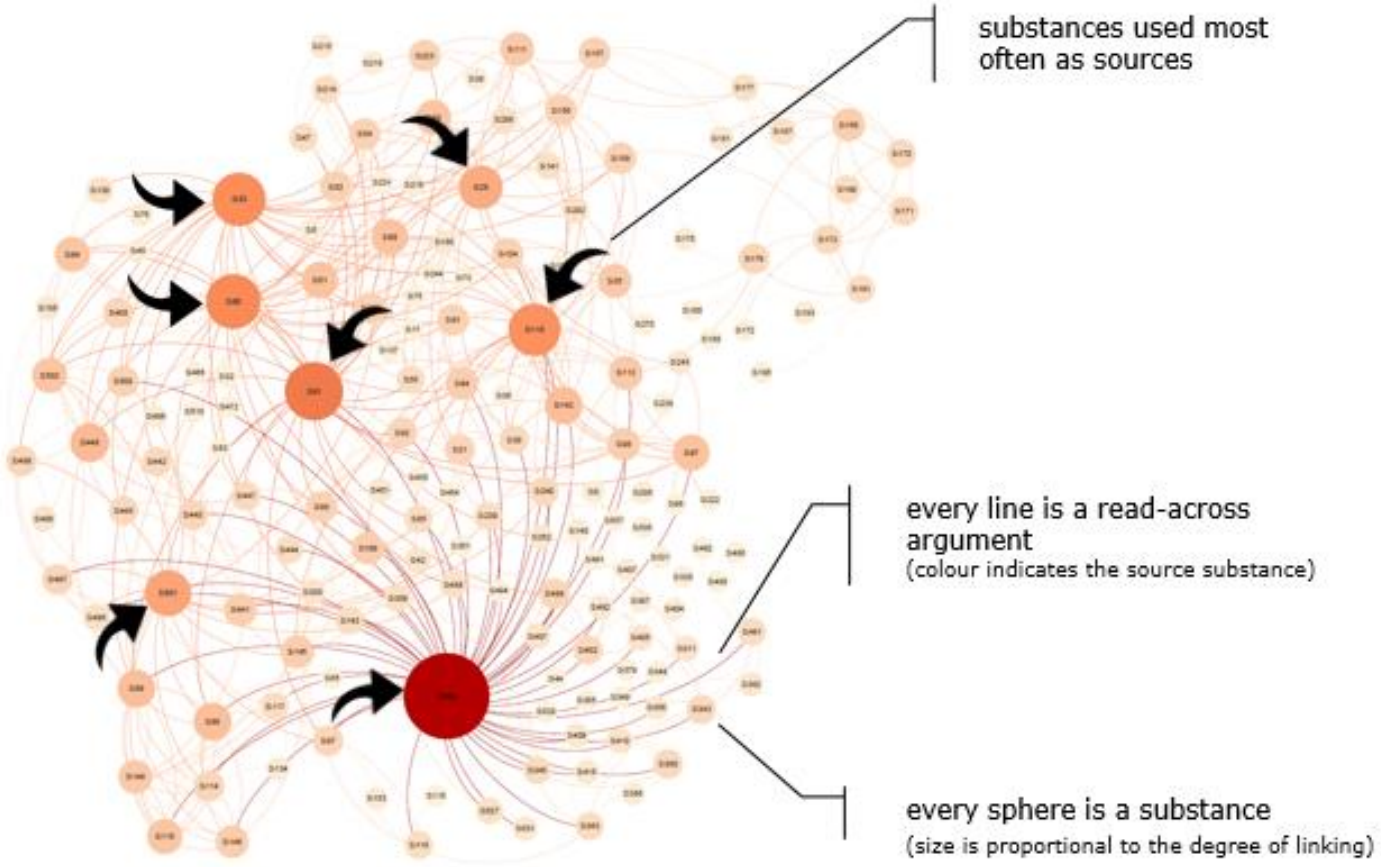
Grouping by read-across/categories

- Grouping is made by collecting analogues from *one-to-one read-across* or *category* statements proposed by either registrants or regulatory authorities

- The following sources of analogues have been used (so far)
  - one-to-one read-across in endpoint study records
  - categories in IUCLID dossiers
  - categories from other international programs (US EPA, IMAP, OECD)

- The list of external sources can be extended further in the future

Visualising substance groups by read-across/categories

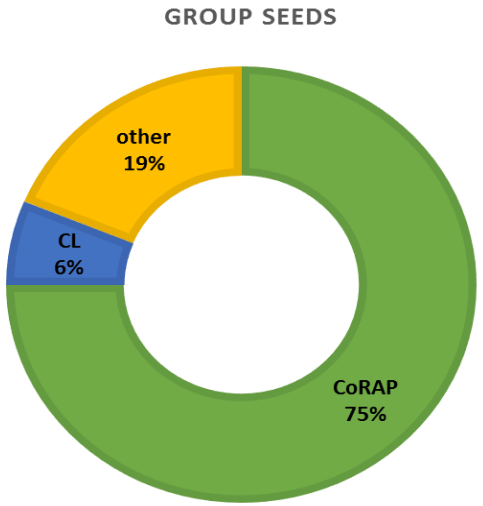


How grouping was applied in round 5 of screening

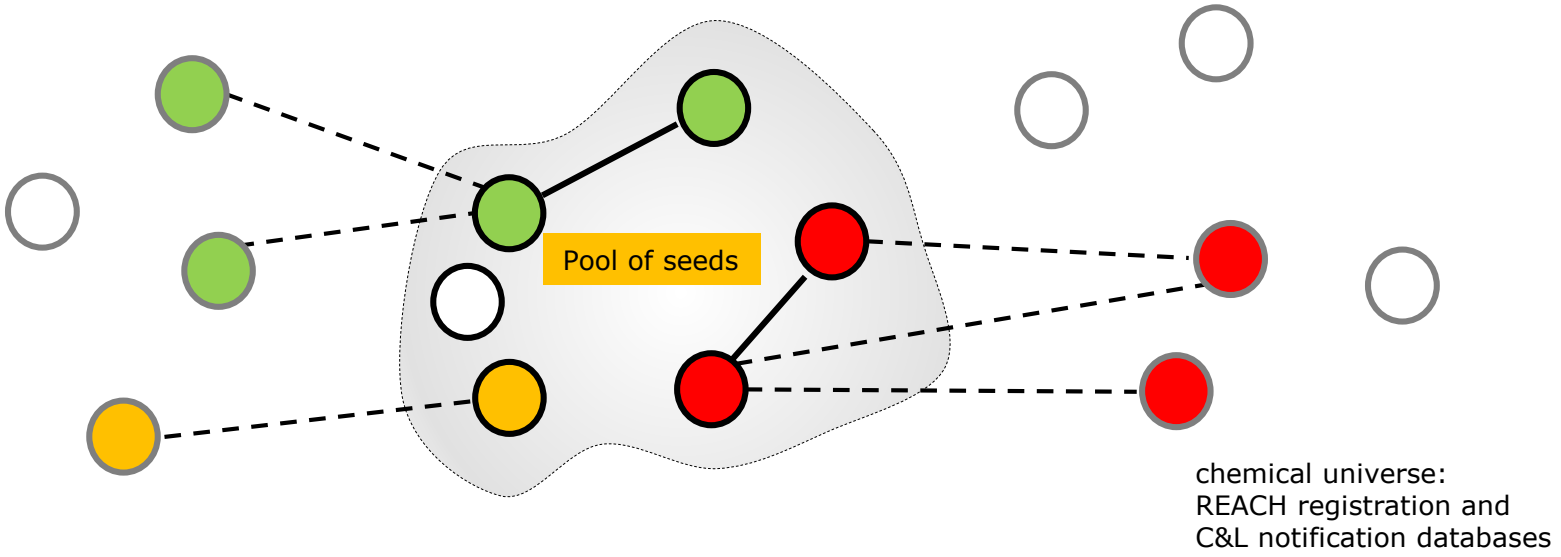
Size of the shortlist and type of entries

- Round 5 shortlist contains **236 substances** of which:
  - **218 belong to 40 groups** and
  - **18** are **individual substances**

- Most of the groups formed around **CoRAP** and **Candidate List 'seeds'**
- Other seeds: candidates from MSCAs and different international regulatory programs (e.g. US EPA TSCA)
- In addition to REACH registered substances, the shortlist also contains substances only notified to the C&L Inventory.



Building of groups in the round5 shortlist



Starting point: selection of substances of suspected/established concern
→ 'seeds' (e.g. CoRAP, CL)

1st step: identification of chemical space we are pulling analogues from
→ REACH registration and C&L notification databases

2nd step: grouping approach between seeds and analogues from identified chemical space
methodology: read-across/categories and structural similarity

Manual check of IT-selected group

- The algorithms occasionally make undesirable linkages between substances
  - the main reason is read-across between structurally different substances (e.g. wrong identifiers of test material, read-across to both cation and anion in case of salts)
  - › Some manual checking of the grouping quality is necessary

- The groups that were considered for the Round 5 short list have been manually checked
  - the vast majority were accepted with no need for manual corrections
  - for a small number of groups some substances were manually removed as their participation in the group seemed erroneous

- **The plausibility of a read-across justification is not evaluated at the group check stage!**

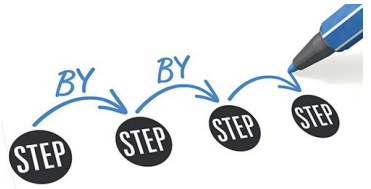
How are groups looked at by MSCAs and tips for the Registrants

How are groups handled by MSCAs?

- The validity of the group and its boundaries may be changed during manual screening
  - for example, the substances that have been related due to structural similarity will be scrutinised to assess if the hazard properties can be different despite the small structural distance
  - additional substances may be added to the group or the group may be split if the grouping approach is refined based on an enhanced understanding on the properties of the substances as manual screening progresses

- MSCAs may propose
  - that the same regulatory process is suitable for all substances in the group
  - different outcomes for different substances in the group

  → *just because the substances are grouped in screening does NOT necessarily mean they will be handled as a group in subsequent regulatory steps*
  - to postpone the assessment of some substances in the group to a later point in time, and after the generation of information for the remaining group members

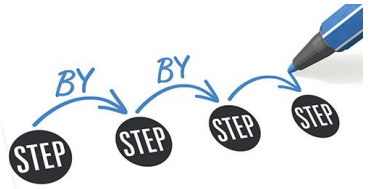


What can registrants do?/1

- Read-across and categories are the most commonly used alternative approach to fulfil the information requirements

- The read-across and category arguments are used at face value by algorithms
  - if the quality of the read-across is poor we may pull together datasets of substances that do no behave similarly
  - when we associate substances we also pull together the hazard findings, that include external experimental data and predictions
  - hence, inclusion of unjustified read-across/category arguments do not necessarily make a stronger case
  - instead they may lead to the identification of additional and perhaps erroneous hazards that need to be followed with the registrant

Read-across assessment framework
https://echa.europa.eu/support/registration/how-to-avoid-unnecessary-testing-on-animals/grouping-of-substances-and-read-across



What can registrants do?/2

- make sure that the identity of all substances used is clear to avoid unintended substance associations
- use read-across and category arguments wisely and adequately and appropriately document them
- explain how structural similarity and dissimilarity affect the predictions
- toxicokinetic information can considerably strengthen the robustness of the read-across
- unsubstantiated arguments of the type "substances are similarly metabolised" are not sufficient to justify the read-across but they trigger our algorithms to pull together hazard datasets
- ECHA's Read-across Assessment Framework structures the scientific evaluation of grouping and read-across in REACH (RAAF)
  - in case the assessment conclusion of your read-across is negative, you may want to re-examine the usefulness of the read-across

Read-across assessment framework
https://echa.europa.eu/support/registration/how-to-avoid-unnecessary-testing-on-animals/grouping-of-substances-and-read-across

Conclusions

- Screening is evolving and Authorities' work is shifting towards groups of related substances
- Methodology of forming groups is improving and ensures fairness to registrants of related substances
- Grouping approach used are mainly *generic* and suitable for *large collections* of substances → verification needed during manual screening by MSCAs!
- Use read-across and category arguments wisely and adequately and appropriately document them

Thank you

giovanni.bernasconi(at)echa.europa.eu

Subscribe to our news at
echa.europa.eu/subscribe

Follow us on Twitter
@EU_ECHA

Follow us on Facebook
Facebook.com/EUECHA