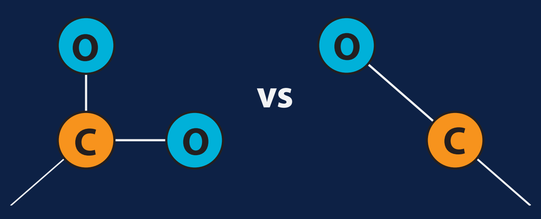
matchbOx is a software implementation of a new chemical structure matching application. Two input structures are provided, as CIFs or SMILES strings, and after matching has occurred, they are rendered inside a 3D model viewer. Fragments of equivalent connectivity and elemental composition within the two structures are colourfully identified, allowing the chemist to rapidly recognise similarities between the two structures.

Part of **Age Concern**, a joint project between **Oxford** and **Durham** Universities

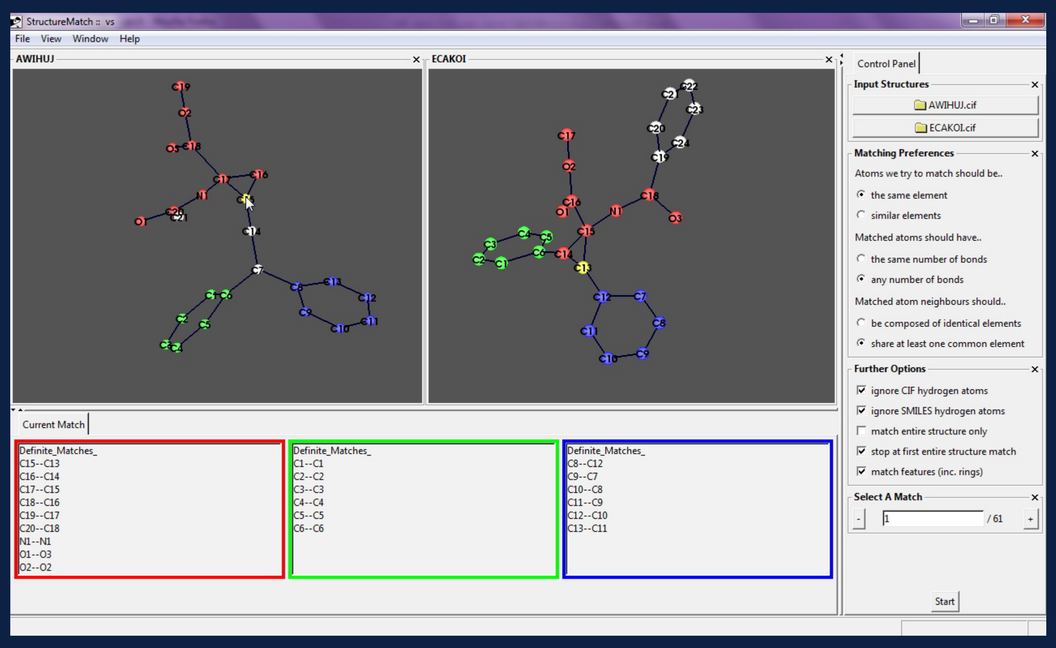
N David Brown | Mustapha Sadki | Amber L Thompson

James Haestier | David J Watkin | *..the Oxford team!*

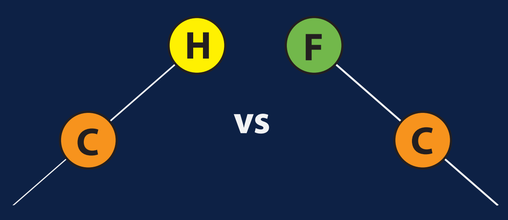
Graph Similarity



Our algorithm builds on an expansion *[1] of the classic Ullmann algorithm*[2]. By the nature of Ullmann's method, connectivity is maintained during structure matching, i.e. there must be the same bonding configuration between the matched atoms in one structure and their counterpart mapped atoms in the other structure. Additionally, we allow tailoring of this graph comparison to successfully match only 'graph identical' environments (i.e. same number of bonds), or 'graph similar' environments where the definition of similarity is arbitrary to the implementation.

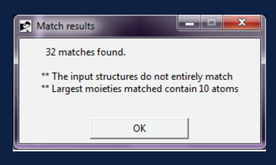


Chemical Similarity



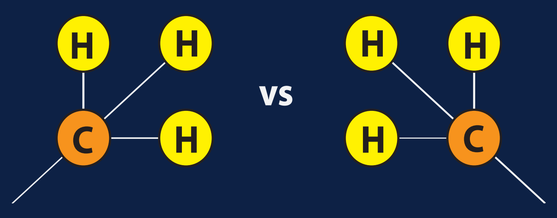
We can tailor which elements will match with which in a similar manner to tailoring our graph matching behaviour. By specifying that only 'chemical identical' candidates and environments should be considered, we limit mapping candidates to those atom pairs where both atoms - one from each input structure - have the same element type. Alternatively, specifying 'chemical similar' for candidates and environments, along with text-described sets of 'similar' environments, allows us to consider equivalent those environments with similar pharmacological effects, for example.

Through an intuitive graphical user interface, users can easily tailor the algorithm's behaviour according to their needs. Elements which should be considered similar may be specified, and the relationship between matched atom bond counts can be defined.

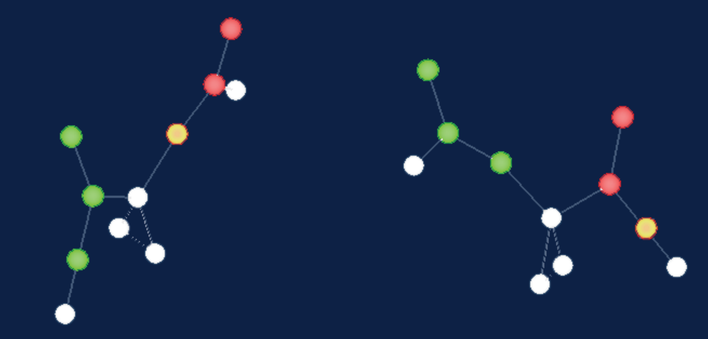


Results are clearly specified in concise text reports.

Arbitrary Mappings



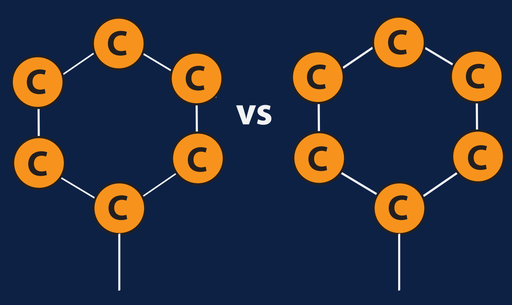
Our algorithm purposefully does not consider any spatial information during its search, and instead relies solely on connectivity to match components of our two structures. One consequence of this is that entirely equivalent matches will be generated between the environments of two matched atoms, if one environment contains more than one atom which is chemically identical or similar to the atoms in the first environment. We handle these so-called 'arbitrary mappings' by clustering the arbitrary atoms in each structure, and associating a cluster in one structure with the relevant cluster in the other; we then know that all mappings where each atom in one cluster maps to one of the atoms in the other cluster are valid.



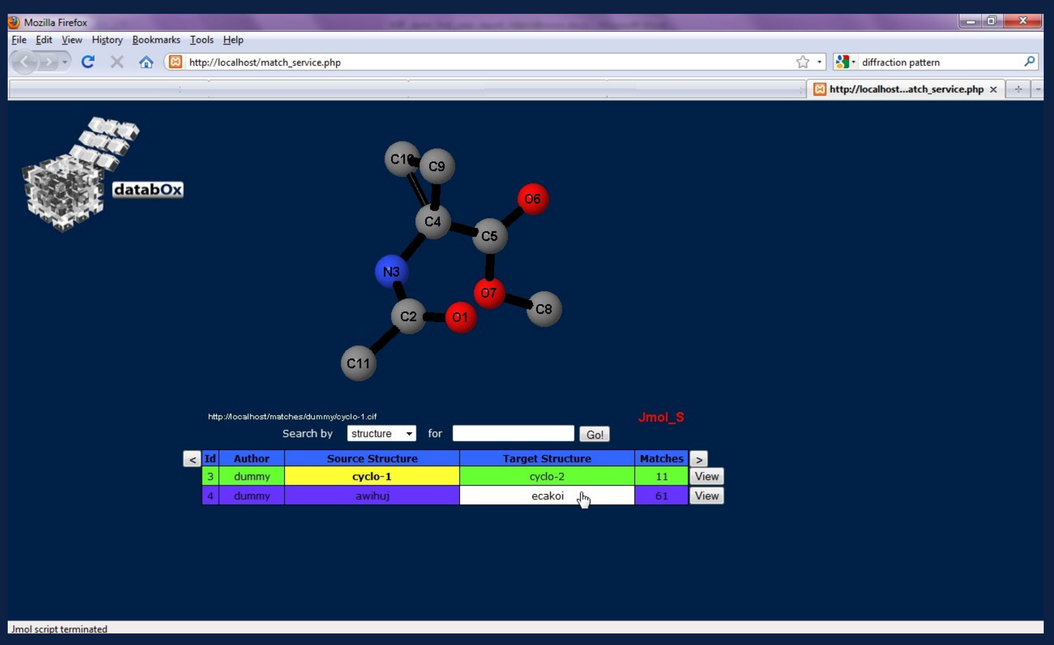
When viewing matches, hovering over a coloured atom in one structure with the mouse will highlight it, along with its matched counterpart in the other structure. Match descriptive text tables can have entries selected with a click, and the relevant matched atom pair will be highlighted in the displays.

Since SMILES strings contain no positional data, a force directed layout strategy is utilised to provide a realistic 3D configuration for the represented structures.

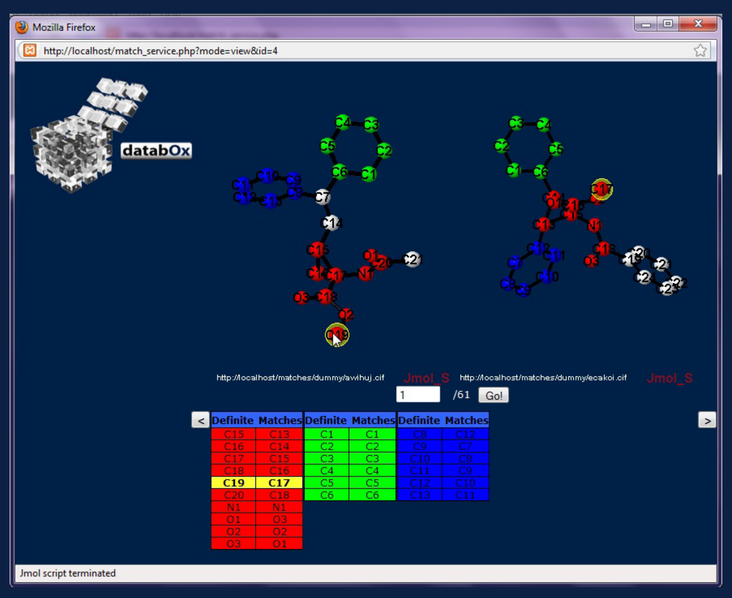
Feature Detection



Certain moieties which exhibit high local symmetry may prolong chemical structure matching, since many mapping permutations may be possible between two such equivalent moieties in the input structures. One related problem is that of 'ring shuffling' - a linear carbon chain in one structure will repeatedly match with a carbon ring (e.g. benzene) in the other structure, shuffling round the ring by one atom to generate each new match. Our implementation is currently able to rapidly detect all rings present in the input structures., as well as any other features specified by the user. This allows more efficient structure matching by avoiding complications such as ring shuffling, and guarantees that any features found in one structure will only match with an equivalent feature in the other structure. It also guarantees that a feature is either matched in its entirety, or else not matched at all.



Match results are communicated to a server, where a continuously running match service called databOx stores records in a database. The service provides a web interface which allows navigation of database entries.



Match records are displayed in a separate viewing window, and allow the same mouse interaction as matchbOx, both with the 3D structure renders and with the match descriptive tables.

References:

1. **L.P. Cordella, P. Foggia, C. Sansone, M. Vento (2004). IEEE Trans. On Patt. Anal. & Mach. Intell., 26, 10, 1367-1372**
2. **J.R. Ullmann (1976). Journal of the Association for Computing Machinery, 23, 31-42 .**