

# Surfing the Organic Chemistry Hyperdocument with CrossFire plus Reactions<sup>1</sup>

Martin G. Hicks

Beilstein-Institut, Varrentrappstrasse 40-42, 60486 Frankfurt, Germany

Received June 30, 1996<sup>®</sup>

The CrossFire system has been extended to permit reaction substructure searching and to give increased performance and access to the Beilstein database. The system is now comprised of three superdocuments: substances (structures and properties), reactions, and citations. These together form the organic chemistry hyperdocument, where the individual documents are hyperlinked to each other. This construction allows the user easy navigation through the organic chemistry hyperdocument by means of data surfing. An example of a reaction substructure search applicable to combinatorial chemistry is shown.

Previous versions of CrossFire had the same basic construction as the Beilstein Handbook or the Beilstein Online database, i.e., a substance orientated database, where the basic unit is an organic compound stored as a registered structure graph together with its corresponding factual data. Under CrossFire, substructure searches and property searches can be carried out, allowing chemical pathways and structure–property relationships to be examined. The CrossFire plus Reactions system requires, in addition to the substance database, a reaction database and the integration of these components into a coherent system. The new system design has three distinct document types: Substances (structures and properties), Reactions, and Citations. The Beilstein Database, loaded under this system, contains currently ca. 7 million substances, 10 million reactions (5 million searchable as reaction substructures) and 20 million chemical, physical, and physiological properties.

**Document Supersymmetry.** There exists a supersymmetry between the documents. Thus a substance document cannot exist without a structure, property, and citation. Similarly a reaction document cannot exist without structures and citations, and a citation document requires either a substance or a reaction document to exist. This supersymmetry has been used to create hyperlinks between the document types.

**Superdocument.** A superdocument is defined as a set of documents of a particular type (or context)—Substances, Reactions, and Citations. It is these documents that are viewed as part of a retrieved hit set by the user. A hit set can be an entire superdocument or (more likely) a subset of it.

**Hyperdocument.** This defines the complete system comprising the three complete superdocuments of substances, reactions, and citations. The hyperdocument forms the information space within which a user navigates.

**Search System.** The CrossFire search and retrieval system has been modified not only to increase performance but also to allow, storage, search, and retrieval of reactions. Reaction substructure searching is now possible with the following attributes: molecule role (reactant or product), reaction centers, bond fate, and atom–atom mapping. The setting of the atom–atom mappings has been solved within the interface using a unique pseudobond approach.

**Interface.** A new user interface was required to give the user the best possible access to the new CrossFire system. The interface was designed to be open in nature and usable with all Beilstein products (CrossFire, Autonom, etc.) and also to allow the use of other structure editors. Thus the Beilstein Commander was developed to give this flexibility and provide efficient and easy access to the database.

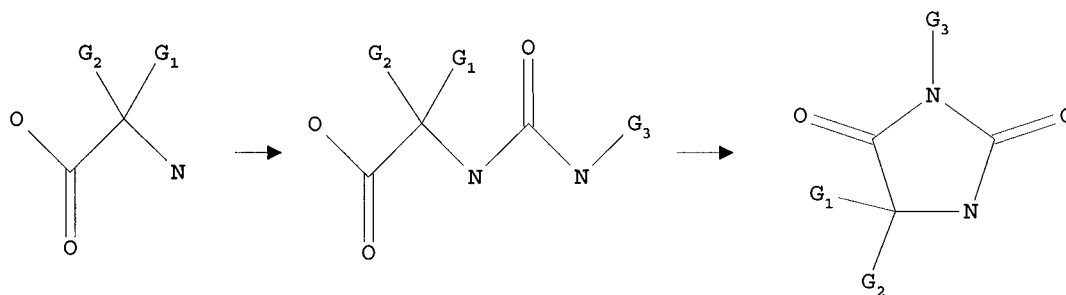
**Hyperlinks.** In each display context (substance, reaction, or citation) there are hyperlinks to the other document types. Thus in a substance display, hyperlinks to the reaction document and to the citation will be present. Similarly in a reaction display, hyperlinks to the substances in the reaction and to the citation will be present, and, also in a citation display, hyperlinks to the substances and reactions in a particular article will be present. Simply clicking on a hyperlink calls up the record in the new context. The number of hyperlinks has been estimated as being ca. 100 million, which compares very favorably with the current estimation of the number of hyperlinks on the World Wide Web, of between 500–1000 million.

**Context Change.** Hyperlinks are a node-to-node link and when invoked are able to display only a single record. To meet the user requirement that it should be possible not only to display a particular citation reference (for example) of a substance property but also to have all of the referenced citations in a hit list, the context change has been developed. This allows the user to convert one superdocument type into another; thus a list of substances can be converted into a list of citations, or a list of reactions can be converted into a list of substances and so on.

**Searching Involving Change of Context.** The ability to change the display context is not only a function that the user can use explicitly but it also takes place in the background when certain combined searches are carried out. It is often of interest to combine searches from two or more different contexts, e.g., a hitlist of substances with that of an author or publication year. It is not technically possible to combine these directly. To enable this to be done, a context change is carried out on the second and any following hit sets to ensure that they are in the same context as the first hit set.

**Data Surfing.** The construction of the CrossFire plus Reactions system, with the Beilstein database, allows the user unrivalled access to organic chemistry information. This access is not only from the point of view of carrying out

<sup>®</sup> Abstract published in *Advance ACS Abstracts*, November 15, 1996.



G1=H, phenyl  
 G2=H, methyl, ethyl, s-butyl, i-butyl, i-propyl, benzyl, phenyl  
 G3=H, butyl, allyl, 2-trifluorotolyl, 4-methoxyphenyl

Figure 1. Synthesis scheme of Hydantoins.

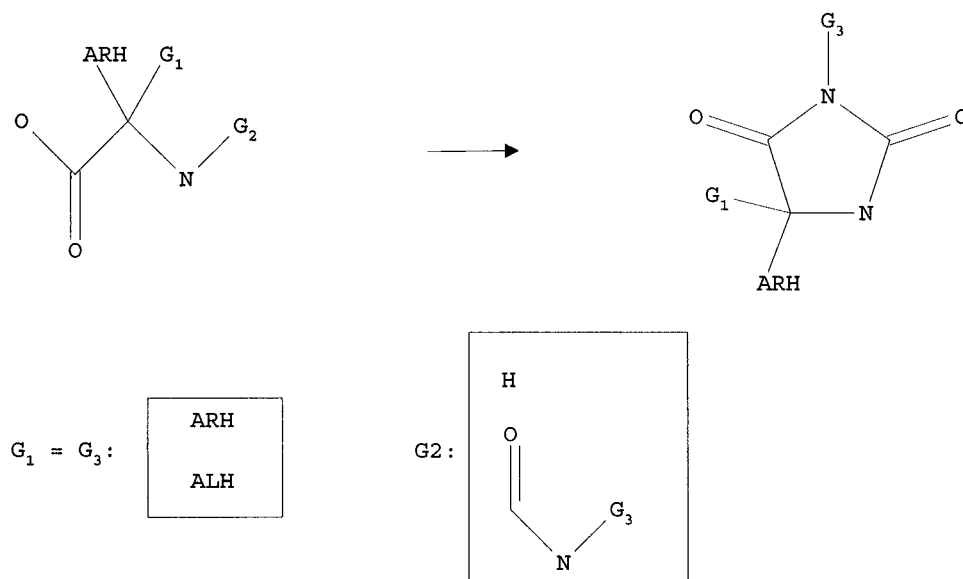


Figure 2. The multistep reaction substructure search query.

straightforward searches but also from the ability to navigate through the organic chemistry hyperdocument using hyperlinks and combinations of structure, reaction, property, and citation searching. Easy navigation allows the user to follow his/her train of thought and retrieve answers in real time. The similarities between this and user behavior on the WWW have given the idea for the definition of this mechanism—*data surfing*. Data surfing allows spontaneity and hence promotes creativity.

**Synthesis Paths.** A synthesis path, from substance to substance, is achieved by clicking the hyperlink to the precursor or product (depending on direction), in the reaction record of the substance display. This allows nondocument-bound synthesis paths to be built up.

**Multistep Reaction Substructure Searching.** The reaction database of ca. 5 million reactions is based on single step reactions which are either individual reactions or the individual steps of a synthesis scheme as published in the original literature. When carrying out reaction substructure searching, it is often not clear how best to search for a particular type of reaction—are all intermediates isolated or is this a one-pot reaction? Clearly, the user can search independently for all possibilities. A simpler alternative is

to carry out the search using a combination of predefined generic groups (ALK = alkyl, etc.) and Markush structures. This is a very powerful tool for many synthesis schemes (but does not work for all reaction types). The use of predefined generic groups is also highly applicable when searching for reactions either used in or relevant to combinatorial chemistry. This is illustrated in the following example: search for synthesis of hydantoins from either an amine or from a urea (result of reacting the amine with an isocyanate) (Figure 1). This was taken from an example of solid phase synthesis. In the example in Figure 2, the amine or urea function is defined by G2. The predefined generic groups ALH (alkyl or hydrogen and ARH (aryl or hydrogen) relax the definition of the query slightly from the original but provide a much easier input, whilst maintaining the general substitution pattern. This search is very fast, taking just over 15 s to give 76 hits in the BS9601 database.

## REFERENCES AND NOTES

- (1) This paper was presented as a product review at the 4th International Conference on Chemical Structures, June 6–10th 1996, Noordwijkerhout, The Netherlands.

CI960111A